(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 November 2004 (11.11.2004)

PCT

(10) International Publication Number
WO 2004/097577 A2

(51) International Patent Classification[7]: G06F

(21) International Application Number: PCT/US2004/012921

(22) International Filing Date: 23 April 2004 (23.04.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/464,983 24 April 2003 (24.04.2003) US

(71) Applicant *(for all designated States except US)*: **NEW YORK UNIVERSITY** [US/US]; 550 First Avenue, New York, NY 10016 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **CHEREPINSKY, Vera** [US/US]; 68 Hope Street, Unit 11, Stamford, CT 06906 (US). **FENG, Jia-Wu** [CN/US]; Box 92, 1230 York Avenue, New York, NY 10021 (US). **REJALI, Marc** [GB/GB]; 57 Danecroft Road, London, SE24 9PA (GB). **MISHRA, Bhubaneswar** [IN/US]; 16 Dunster Road, Great Neck, NY 11021 (US).

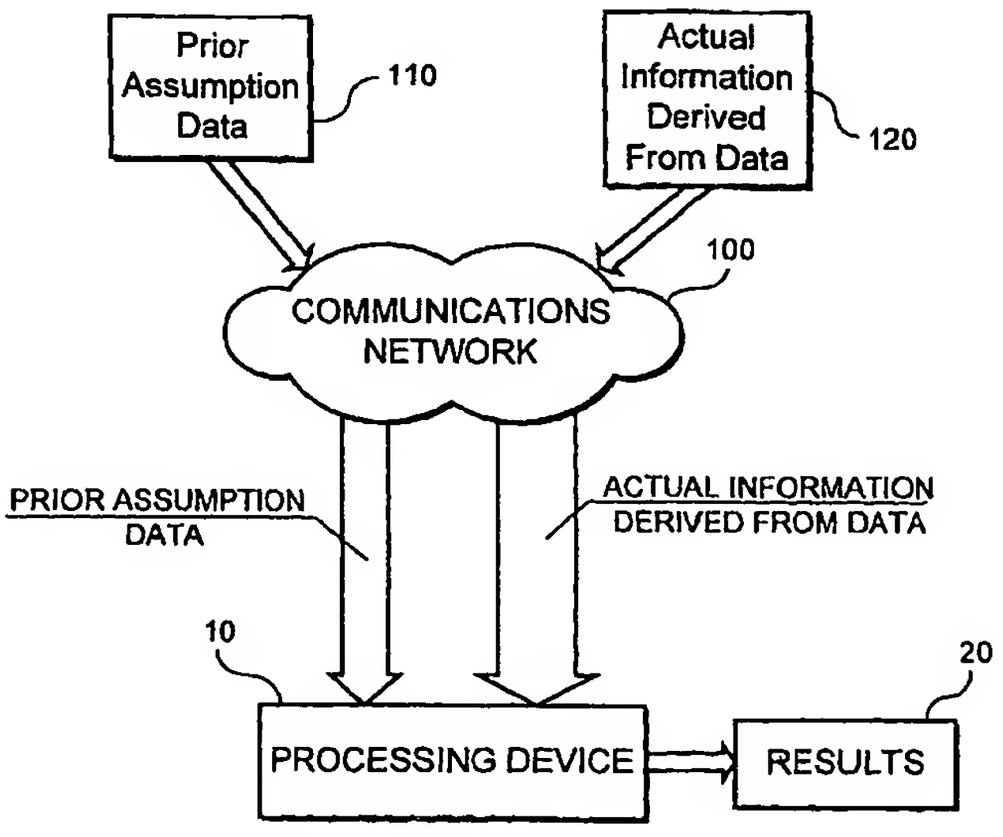
(74) Agent: **ABELEV, Gary**; Baker Botts L.L.P., 30 Rockefeller Plaza, New York, NY 10112-4498 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHODS, SOFTWARE ARRANGEMENTS, STORAGE MEDIA, AND SYSTEMS FOR PROVIDING A SHRINKAGE-BASED SIMILARITY METRIC



(57) Abstract: The present invention relates to systems, methods, and software arrangements for determining associations between two or more datasets. The systems, methods, and software arrangements used to determine such associations include a determination of a correlation coefficient that incorporates both prior assumptions regarding such datasets and actual information regarding the datasets. The systems, methods, and software arrangements of the present invention can be useful in an analysis of microarray data, including gene expression arrays, to determine correlations between genotypes and phenotypes. Accordingly, the systems, methods, and software arrangements of the present invention may be utilized to determine a genetic basis of complex genetic disorder (e.g. those characterized by the involvement of more than one gene).

WO 2004/097577 A2

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHODS, SOFTWARE ARRANGEMENTS, STORAGE MEDIA, AND SYSTEMS FOR PROVIDING A SHRINKAGE-BASED SIMILARITY METRIC

CROSS REFERENCE TO RELATED APPLICATION

This application claims priority from U.S. Patent Application Serial No. 60/464,983 filed on April 24, 2003, the entire disclosure of which is incorporated herein by reference.

FIELD OF THE INVENTION

The present invention relates generally to systems, methods, and software arrangements for determining associations between one or more elements contained within two or more datasets. For example, the embodiments of systems, methods, and software arrangements determining such associations may obtain a correlation coefficient that incorporates both prior assumptions regarding two or more datasets and actual information regarding such datasets.

BACKGROUND OF THE INVENTION

Recent improvements in observational and experimental techniques allow those of ordinary skill in the art to better understand the structure of a substantially unobservable transparent cell. For example, microarray-based gene expression analysis may allow those of ordinary skill in the art to quantify the transcriptional states of cells. Partitioning or clustering genes into closely related groups has become an important mathematical process in the statistical analyses of microarray data.

Traditionally, algorithms for cluster analysis of genome-wide expression data from DNA microarray hybridization were based upon statistical properties of gene expressions, and result in organizing genes according to similarity in pattern of gene expression. These algorithms display the output graphically, often in a binary tree form, conveying the clustering and the underlying expression data simultaneously. If two genes belong to the same cluster (or, equivalently, if they

belong to the same subtree of small depth), then it may be possible to infer a common regulatory mechanism for the two genes, or to interpret this information as an indication of the status of cellular processes. Furthermore, a coexpression of genes of known function with novel genes may result in a discovery process for characterizing unknown or poorly characterized genes. In general, false negatives (where two coexpressed genes are assigned to distinct clusters) may cause the discovery process to ignore useful information for certain novel genes, and false positives (where two independent genes are assigned to the same cluster) may result in noise in the information provided to the subsequent algorithms used in analyzing regulatory patterns. Consequently, it may be important that the statistical algorithms for clustering are reasonably robust. Nevertheless, the microarray experiments that can be carried out in an academic laboratory at a reasonable cost are minimal, and suffer from an experimental noise. As such, those of ordinary skill in the are may use certain algorithms to deal with small sample data.

One conventional clustering algorithm is described in Eisen *et al.* ("Eisen"), *Proc. Natl. Acad. Sci. USA* 95, 14863-14868 (1998). In Eisen, the gene-expression data were collected on spotted DNA microarrays (*See, e.g.*, Schena *et al.* ("Schena"), *Proc. Natl. Acad. Sci. USA* 93, 10614-10619 (1996)), and were based upon gene expression in the budding yeast *Saccharomyces cerevisiae* during the diauxic shift (*See, e.g.*, DeRisi *et al.* ("DeRisi"), *Science* 278, 680-686 (1997)), the mitotic cell division cycle (*See, e.g.*, Spellman *et al.* ("Spellman"), *Mol. Biol. Cell* 9, 3273-3297 (1998)), sporulation (*See, e.g.*, Chu *et al.* ("Chu"), *Science* 282, 699-705 (1998)), and temperature and reducing shocks. The disclosures of each of these references are incorporated herein by reference in their entireties. In Eisen, RNA from experimental samples (taken at selected times during the process) were labeled during reverse transcription with a red-fluorescent dye Cy5, and mixed with a reference sample labeled in parallel with a green-fluorescent dye Cy3. After hybridization and appropriate washing steps, separate images were acquired for each fluorophor, and fluorescence intensity ratios obtained for all target elements. The experimental data were provided in an *MxN* matrix structure, in which the *M* rows represented all genes for which data had been collected, the *N* columns represented individual array experiments (*e.g.*, single time points or conditions), and each entry

represented the measured Cy5/Cy3 fluorescence ratio at the corresponding target element on the appropriate array. All ratio values were log-transformed to treat inductions and repressions of identical magnitude as numerically equal but opposite in sign. In Eisen, it was assumed that the raw ratio values followed log-normal distributions and hence, the log-transformed data followed normal distributions.

The gene similarity metric employed in this publication was a form of a correlation coefficient. Let G_i be the (log-transformed) primary data for a gene G in condition i. For any two genes X and Y observed over a series of N conditions, the classical similarity score based upon a Pearson correlation coefficient is:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right),$$

where

$$\Phi_G{}^2 = \sum_{i=1}^{N}\frac{(G_i - G_{offset})^2}{N}$$

and G_{offset} is the estimated mean of the observations, *i.e.*,

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{i=1}^{N} G_i.$$

Φ_G is the (rescaled) estimated standard deviation of the observations. In the Pearson correlation coefficient model, G_{offset} is set equal to 0. Nevertheless, in the analysis described in Eisen, "values of G_{offset} which are not the average over observations on G were used when there was an assumed unchanged or reference state represented by the value of G_{offset}, against which changes were to be analyzed; in all of the examples presented there, G_{offset} was set to 0, corresponding to a fluorescence ratio of 1.0." To

distinguish this modified correlation coefficient from the classical Pearson correlation coefficient, we shall refer to it as Eisen correlation coefficient. Nevertheless, setting G_{offset} equal to 0 or 1 results in an increase in false positives or false negatives, respectively.

SUMMARY OF THE INVENTION

The present invention relates generally to systems, methods, and software arrangements for determining associations between one or more elements contained within two or more datasets. An exemplary embodiment of the systems, methods, and software arrangements determining the associations may obtain a correlation coefficient that incorporates both prior assumptions regarding two or more datasets and actual information regarding such datasets. For example, an exemplary embodiment of the present invention is directed toward systems, methods, and software arrangements in which one of the prior assumptions used to calculate the correlation coefficient is that an expression vector mean μ of each of the two or more datasets is a zero-mean normal random variable (with an *a priori* distribution $N(0,\tau^2)$), and in which one of the actual pieces of information is an *a posteriori* distribution of expression vector mean μ that can be obtained directly from the data contained in the two or more datasets. The exemplary embodiment of the systems, methods, and software arrangements of the present invention are more beneficial in comparison to conventional methods in that they likely produce fewer false negative and/or false positive results. The exemplary embodiment of the systems, methods, and software arrangements of the present invention are further useful in the analysis of microarray data (including gene expression arrays) to determine correlations between genotypes and phenotypes. Thus, the exemplary embodiments of the systems, methods, and software arrangements of the present invention are useful in elucidating the genetic basis of complex genetic disorders (*e.g.*, those characterized by the involvement of more than one gene).

According to the exemplary embodiment of the present invention, a similarity metric for determining an association between two or more datasets may take the form of a correlation coefficient. However, unlike conventional correlations, the correlation coefficient according to the exemplary embodiment of the present

invention may be derived from both prior assumptions regarding the datasets (including but not limited to the assumption that each dataset has a zero mean), and actual information regarding the datasets (including but not limited to an *a posteriori* distribution of the mean). Thus, in one the exemplary embodiment of the present invention, a correlation coefficient may be provided, the mathematical derivation of which can be based on James-Stein shrinkage estimators. In this manner, it can be ascertained how a shrinkage parameter of this correlation coefficient may be optimized from a Bayesian point of view, *e.g.*, by moving from a value obtained from a given dataset toward a "believed" or theoretical value. For example, in one exemplary embodiment of the present invention, G_{offset} of the gene similarity metric described above may be set equal to $\gamma\overline{G}$, where γ is a value between 0.0 and 1.0. When $\gamma = 1.0$, the resulting similarity metric may be the same as the Pearson correlation coefficient, and when $\gamma = 0.0$, it may be the same as the Eisen correlation coefficient. However, for a non-integer value of γ (*i.e.*, a value other than 0.0 or 1.0), the estimator for $G_{offset} = \gamma\overline{G}$ can be considered as the unbiased estimator $\overline{G}$ decreasing toward the believed value for G_{offset}. This optimization of the correlation coefficient can minimize the occurrence of false positives relative to the Eisen correlation coefficient, and the occurrence of false negatives relative to the Pearson correlation coefficient.

According to an exemplary embodiment of the present invention, the general form of the following equation:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right), \quad (1)$$

where

$$\Phi_G = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - G_{offset})^2} \quad \text{and}$$
$$G_{offset} = \gamma\bar{G} \quad \text{for} \quad G \in \{X, Y\} \tag{2}$$

can be used to derive a similarity metric which is dictated by the data. In a general setting, all values X_{ij} for gene j may have a Normal distribution with mean θ_j and standard deviation β_j (variance β_j^2); *i.e.*, $X_{ij} \sim N(\theta_j, \beta_j^2)$ for $i = 1,\ldots,N$, with j fixed ($1 \leq j \leq M$), where θ_j is an unknown parameter (taking different values for different j). For the purpose of estimation, θ_j can be assumed to be a random variable taking values close to zero: $\theta_j \sim N(0, \tau^2)$.

In one exemplary embodiment of the present invention, the posterior distribution of θ_j may be derived from the prior $N(0,\tau^2)$ and the data via the application of James-Stein Shrinkage estimators. θ_j then may be estimated by its mean. In another exemplary embodiment, the James-Stein Shrinkage estimators are W and $\hat{\beta}^2$.

In yet another exemplary embodiment of the present invention, the posterior distribution of θ_j may be derived from the prior $N(0,\tau^2)$ and the data from the Bayesian considerations. θ_j then may be estimated by its mean.

The present invention further provides exemplary embodiments of the systems, methods, and software arrangements for implementation of hierarchical clustering of two or more datapoints in a dataset. In one preferred embodiment of the present invention, the datapoints to be clustered can be gene expression levels obtained from one or more experiments, in which gene expression levels may be analyzed under two or more conditions. Such data documenting alterations in the gene expression under various conditions may be obtained by microarray-based genomic analysis or other high-throughput methods known to those of ordinary skill in the art. Such data may reflect the changes in gene expression that occur in response to alterations in various phenotypic indicia, which may include but are not limited to developmental and/or pathophysiological (*i.e.*, disease-related) changes. Thus, in one exemplary embodiment of the present invention, the establishment of

genotype/phenotype correlations may be permitted. The exemplary systems, methods, and software arrangements of the present invention may also obtain genotype/phenotype correlations in complex genetic disorders, *i.e.*, those in which more than one gene may play a significant role. Such disorders include, but are not limited to, cancer, neurological diseases, developmental disorders, neurodevelopmental disorders, cardiovascular diseases, metabolic diseases, immunologic disorders, infectious diseases, and endocrine disorders.

According to still another exemplary embodiment of the present invention, a hierarchical clustering pseudocode may be used in which a clustering procedure is utilized by selecting the most similar pair of elements, starting with genes at the bottom-most level, and combining them to create a new element. In one exemplary embodiment of the present invention, the "expression vector" for the new element can be the weighted average exemplary of the expression vectors of the two most similar elements that were combined. In another embodiment of the present invention, the structure of repeated pair-wise combinations may be represented in a binary tree, whose leaves can be the set of genes, and whose internal nodes can be the elements constructed from the two children nodes.

In another preferred embodiment of the present invention, the datapoints to be clustered may be values of stocks from one or more stock markets obtained at one or more time periods. Thus, in this preferred embodiment, the identification of stocks or groups of stocks that behave in a coordinated fashion relative to other groups of stocks or to the market as a whole can be ascertained. The exemplary embodiment of the systems, methods, and software arrangements of the present invention therefore may be used for financial investment and related activities.

For a better understanding of the present invention, together with other and further objects, reference is made to the following description, taken in conjunction with the accompanying drawings, and its scope will be pointed out in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

For a more complete understanding of the present invention and its advantages, reference is now made to the following description, taken in conjunction with the accompanying drawings, in which:

Figure 1 is a first exemplary embodiment of a system according to the present invention for determining an association between two datasets based on a combination of data regarding one or more prior assumptions about the datasets and actual information derived from such datasets;

Figure 2 is a second exemplary embodiment of the system according to the present invention for determining the association between the datasets;

Figure 3 is an exemplary embodiment of a process according to the present invention for determining the association between two datasets which can utilize the exemplary systems of Figures 1 and 2;

Figure 4 is an exemplary illustration of histograms generated by performing in silico experiments with the four different algorithms, under four different conditions;

Figure 5 is a schematic diagram illustrating the regulation of cell-cycle functions of yeast by various translational activators (Simon *et al.*, *Cell* 106: 67-708 (2001)), used as a reference to test the performance of the present invention;

Figure 6 depicts Receiver Operator Characteristic (ROC) curves for each of the three algorithms *Pearson, Eisen* or *Shrinkage*, in which each curve is parameterized by the cut-off value $\theta \in \{1.0,0.95,...,-1.0\}$;

Figures 7A-B show FN (Panel A) and FP (Panel B) curves, each plotted as a function of θ; and

Figure 8 shows ROC curves, with threshold plotted on the z–axis.

DETAILED DESCRIPTION OF THE INVENTION

An exemplary embodiment of the present invention provides systems, methods, and software arrangements for determining one or more associations between one or more elements contained within two or more datasets. The determination of such associations may be useful, *inter alia*, in ascertaining

coordinated changes in a gene expression that may occur, for example, in response to alterations in various phenotypic indicia, which may include (but are not limited to) developmental and/or pathophysiological (*i.e.*, disease-related) changes establishment of these genotype/phenotype correlations can permit a better understanding of a direct or indirect role that the identified genes may play in the development of these phenotypes. The exemplary systems, methods, and software arrangements of the present invention can further be useful in elucidating genotype/phenotype correlations in complex genetic disorders, *i.e.*, those in which more than one gene may play a significant role. The knowledge concerning these relationships may also assist in facilitating the diagnosis, treatment and prognosis of individuals bearing a given phenotype. The exemplary systems, methods, and software arrangements of the present invention also may be useful for financial planning and investment.

Figure 1 illustrates a first exemplary embodiment of a system for determining one or more associations between one or more elements contained within two or more datasets. In this exemplary embodiment, the system includes a processing device 10 which is connected to a communications network 100 (*e.g.*, the Internet) so that it can receive data regarding prior assumptions about the datasets and/or actual information determined from the datasets. The processing device 10 can be a mini-computer (*e.g.*, Hewlett Packard mini computer), a personal computer (*e.g.*, a Pentium chip-based computer), a mainframe computer (*e.g.*, IBM 3090 system), and the like. The data can be provided from a number of sources. For example, this data can be prior assumption data 110 obtained from theoretical considerations or actual data 120 derived from the dataset. After the processing device 10 receives the prior assumption data 110 and the actual information 120 derived from the dataset via the communications network 100, it can then generate one or more results 20 which can include an association between one or more elements contained in one or more datasets.

Figure 2 illustrates a second exemplary embodiment of the system 10 according to the present invention in which the prior assumption data 110 obtained from theoretical considerations or actual data 120 derived from the dataset is transmitted to the system 10 directly from an external source, *e.g.*, without the use of the communications network 100 for such transfer of the data. In this second

exemplary embodiment of the system 10, it is also possible for the prior assumption data 110 obtained from theoretical considerations or the actual information 120 derived from the dataset to be obtained from a storage device provided in or connected to the processing device 10. Such storage device can be a hard drive, a CD-ROM, *etc.* which are known to those having ordinary skill in the art.

Figure 3 shows an exemplary flow chart of the embodiment of the process according to the present invention for determining an association between two datasets based on a combination of data regarding one or more prior assumptions about and actual information derived from the datasets. This process can be performed by the exemplary processing device 10 which is shown in Figures 1 or 2. As shown in Figure 3, the processing device 10 receives the prior assumption data 110 (first data) obtained from theoretical considerations in step 310. In step 320, the processing device 10 receives actual information 120 derived from the dataset (second data). In step 330, the prior assumption (first) data obtained 110 from theoretical considerations and the actual (second) data 120 derived from the dataset are combined to determine an association between two or more datasets. The results of the association determination are generated in step 340.

I. OVERALL PROCESS DESCRIPTION

The exemplary systems, methods, and software arrangements according to the present invention may be (*e.g.*, as shown in Figures 1-3) used to determine the associations between two or more elements contained in datasets to obtain a correlation coefficient that incorporates both prior assumptions regarding the two or more datasets and actual information regarding such datasets. One exemplary embodiment of the present invention provides a correlation coefficient that can be obtained based on James-Stein Shrinkage estimators, and teaches how a shrinkage parameter of this correlation coefficient may be optimized from a Bayesian point of view, moving from a value obtained from a given dataset toward a "believed" or theoretical value. Thus, in one exemplary embodiment of the present invention, G_{offset} may be set equal to $\gamma\overline{G}$, where γ is a value between 0.0 and 1.0. When $\gamma = 1.0$, the resulting similarity metric γ may be the same as the Pearson correlation coefficient,

and when $\gamma = 0.0$, γ may be the same as the Eisen correlation coefficient. For a non-integer value of γ (*i.e.*, a value other than 0.0 or 1.0), the estimator for $G_{offset} = \gamma\overline{G}$ can be considered as an unbiased estimator $\overline{G}$ decreasing toward the believed value for G_{offset}. Such exemplary optimization of the correlation coefficient may minimize the occurrence of false positives relative to the Eisen correlation coefficient and minimize the occurrence of false negatives relative to the Pearson correlation coefficient.

II. EXEMPLARY MODEL

A family of correlation coefficients parameterized by $0 \leq \gamma \leq 1$ may be defined as follows:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right), \tag{1}$$

where

$$\Phi_G = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - G_{offset})^2} \quad \text{and}$$
$$G_{offset} = \gamma G \quad \text{for} \quad G \in \{X, Y\} \tag{2}$$

In contrast, the Pearson Correlation Coefficient uses $G_{offset} = \overline{G} = \frac{1}{N}\sum_{j=1}^{N} G_i$ for every gene G, or $\gamma = 1$, and the Eisen Correlation Coefficient uses $G_{offset} = 0$ for every gene G, or $\gamma = 0$.

In an exemplary embodiment of the present invention, the general form of equation (1) may be used to derive a similarity metric which is dictated by both the data and prior assumptions regarding the data, and that reduces the occurrence of false

positives (relative to the Eisen metric) and false negatives (relative to the Pearson correlation coefficient).

SETUP

As described above, the metric used by Eisen had the form of equation (1) with G_{offset} set to 0 for every gene G (as a reference state against which to measure the data). Nevertheless, even if it is initially assumed that each gene G has zero mean, such assumption should be updated when data becomes available. In an exemplary embodiment of the present invention, gene expression data may be provided in the form of the levels of M genes expressed under N experimental conditions. The data can be viewed as

$$\{\{X_{ij}\}_{i=1}^{N}\}_{j=1}^{M}$$

where $M \gg N$ and $\{X_{ij}\}_{i=1}^{N}$ is the data vector for gene j.

DERIVATION

S may be rewritten in the following notation:

$$\begin{aligned} S(X_j, X_k) &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij}-(X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik}-(X_k)_{offset}}{\Phi_k}\right), \\ \Phi_j{}^2 &= \frac{1}{N}\sum_{i}\left(X_{ij}-(X_j)_{offset}\right)^2 \end{aligned} \tag{3}$$

In a general setting, the following exemplary assumptions may be made regarding the data distribution: let all values X_{ij} for gene j have a Normal distribution with mean θ_j and standard deviation β_j (variance β_j^2); *i.e.*, $X_{ij} \sim N(\theta_j,\beta_j^2)$ for $i = 1,\ldots,N$, with j fixed

($1 \leq j \leq M$), where θ_j is an unknown parameter (taking different values for different j). For the purpose of estimation, θ_j can be assumed to be a random variable taking values close to zero: $\theta_j \sim N(0, \tau^2)$.

It is also possible according to the present invention to assume that the data are range-normalized, such that $\beta_j^2 = \beta^2$ for every j. If this exemplary assumption does not hold true on a given data set, it can be corrected by scaling each gene vector appropriately. Using conventional methods, the range may be adjusted to scale to an interval of unit length, *i.e.*, its maximum and minimum values differ by 1. Thus, $X_{ij} \sim N(\theta_j, \beta_j^2)$ and $\theta_j \sim N(\theta, \tau^2)$.

Replacing $(X_j)_{offset}$ in equation (3) by the exact value of the mean θ_j may yield a *Clairvoyant* correlation coefficient of X_j and X_k. Nevertheless, because θ_j is a random variable, it should be estimated from the data. Therefore, to obtain an explicit formula for $S(X_j, X_k)$, it is possible to derive estimators $\hat{\theta}_j$ for all j.

In Pearson correlation coefficient, θ_j may be estimated by the vector mean $\overline{X}_{.j}$; and the Eisen correlation coefficient corresponds to replacing θ_j by 0 for every j, which is equivalent to assuming $\theta_j \sim N(0,0)$ (*i.e.*, $\tau^2 = 0$). In an exemplary embodiment of the system, method, and software arrangement according to the present invention, an estimate of θ_j (call it $\hat{\theta}_j$) may be determined that takes into account both the prior assumption and the data.

ESTIMATION OF θ_j

a. N=1

First, it is possible according to the present invention to obtain the posterior distribution of θ_j from the prior $N(0, \tau^2)$ and the data. This exemplary derivation can be done either from the Bayesian considerations, or via the James-Stein Shrinkage estimators (*See, e.g.*, James *et al.* ("James"), *Proc. 4th Berkeley Symp. Math. Statist.* Vol. 1, 361-379 (1961); and Hoffman, *Statistical Papers* 41(2), 127-158 (2000), the disclosures of which are incorporated herein by reference in their entireties). In this exemplary embodiment of the present invention, the Bayesian estimators method can be applied, and it may initially be assumed that $N = 1$, *i.e.*,

there is one data point for each gene. Moreover, the variance can initially be denoted by σ^2, such that:

$$X_j \sim N(\theta_j, \sigma^2) \tag{4}$$

$$\theta_j \sim N(\theta, \tau^2) \tag{5}$$

For the sake of clarity, the probability density function (pdf) of θ_j can be denoted by $\pi(\cdot)$, and the pdf of X_j can be denoted by $f(\cdot)$. Based on equations (4) and (5), the following relationships may be derived:

$$\begin{aligned} \pi(\theta_j) &= \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\theta_j{}^2/2\tau^2\right), \\ f(X_j|\theta_j) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(X_j-\theta_j)^2/2\sigma^2\right). \end{aligned}$$

By Bayes' Rule, the joint pdf of X_j and θ_j may be given by

$$\begin{aligned} f(X_j, \theta_j) &= f(X_j|\theta_j)\ \pi(\theta_j) \\ &= \frac{1}{2\pi\sigma\tau} \exp\left(-\left[\frac{\theta_j{}^2}{2\tau^2} + \frac{(X_j-\theta_j)^2}{2\sigma^2}\right]\right) \end{aligned} \tag{6}$$

Then $f(X_j)$, the marginal pdf of X_j may be

$$\begin{aligned} f(X_j) &= \mathbf{E}_{\theta_j} f(X_j|\theta_j) = \int_{\theta=-\infty}^{\infty} f(X_j|\theta)\pi(\theta)d\theta \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\left(-\frac{X_j{}^2}{2(\sigma^2+\tau^2)}\right), \end{aligned} \tag{7}$$

where the equality in equation (7) is written out in Appendix A.2. Based again on Bayes' Theorem, the posterior distribution of θ_j may be given by:

$$\pi(\theta_j|X_j) = \frac{f(X_j,\theta_j)}{f(X_j)}$$
$$= \frac{f(X_j|\theta_j)\ \pi(\theta_j)}{f(X_j)} \quad \text{by (6)}$$
$$= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\ \exp\left[-\frac{\left(\theta_j - \frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2}{2\left(\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)}\right].$$

(8)

(See Appendix A.3 for derivation of equation (8).)

Since this has a Normal form, it can be determined that:

$$\begin{aligned}\mathbf{E}(\theta_j|X_j) &= \frac{\tau^2}{\sigma^2+\tau^2}X_j \\ &= \left(1-\frac{\sigma^2}{\sigma^2+\tau^2}\right)X_j, \\ Var(\theta_j|X_j) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.\end{aligned}$$

(9)

θ_j then may be estimated by its mean.

b. *N* IS ARBITRARY

In contrast to above where N was selected to be 1, if *N* is selected to be arbitrary and greater than 1, X_j becomes a vector $X_{.j}$. It can be shown using likelihood functions that the vector of values $\{X_{ij}\}_{i=1}^{N}$, with $X_{ij} \sim N(\theta_j, \beta^2)$ may be treated as a single data point $Y_j = \overline{X}_{.j} = \sum_{i=1}^{N} X_{ij}/N$ from the distribution

$N(\theta_j, \beta^2/N)$ (see Appendix A.4). Thus, following the above derivation with $\sigma^2 = \beta^2/N$, a Bayesian estimator for θ_j may be given by $\mathbf{E}(\theta_j|X_{.j})$:

$$\widehat{\theta_j} = \left(1 - \frac{\beta^2/N}{\beta^2/N + \tau^2}\right) Y_j. \tag{10}$$

However, equation (10) may likely not be directly used in equation (3) because τ^2 and β^2 may be unknown, such that τ^2 and β^2 should be estimated from the data.

c. ESTIMATION OF $1/(\beta^2/N+\tau^2)$

In this exemplary embodiment of the present invention, let

$$W = \frac{M-2}{\sum_{j=1}^{M} Y_j^2}. \tag{11}$$

This equation for W is obtained from James-Stein estimation. W may be treated as an educated guess of an estimator for $1/(\beta^2/N+\tau^2)$, and it can be verified that W is an appropriate estimator for $1/(\beta^2/N+\tau^2)$, as follows:

$$\begin{aligned} Y_j &\sim \theta_j + \frac{\beta^2}{N}\mathcal{N}(0,1) \\ &\sim \tau^2\mathcal{N}(0,1) + \frac{\beta^2}{N}\mathcal{N}(0,1) \\ &\sim \left(\frac{\beta^2}{N} + \tau^2\right)\mathcal{N}(0,1) \sim \mathcal{N}\left(0, \frac{\beta^2}{N} + \tau^2\right) \end{aligned} \tag{12}$$

The transition in equation is set forth in Appendix A.5. If we let $\alpha^2=\beta^2/N+\tau^2$, then from equation (12) it follows that:

$$\frac{Y_j}{\sqrt{\alpha^2}} = \frac{Y_j}{\alpha} \sim \mathcal{N}(0,1),$$

and hence

$$\sum_{j=1}^{M} {Y_j}^2 = \alpha^2 \sum_{j=1}^{M} \left(\frac{Y_j}{\alpha}\right)^2 = \alpha^2 \chi^2_M,$$

where X^2_M is a Chi-square random variable with M degrees of freedom. By properties of the Chi-square distribution and the linearity of expectation,

$$\begin{aligned} \mathbf{E}\left(\frac{\alpha^2}{\sum {Y_j}^2}\right) &= \frac{1}{M-2} \text{ (see Appendix A.6)} \\ \mathbf{E}(W) &= \mathbf{E}\left(\frac{M-2}{\sum {Y_j}^2}\right) = \frac{1}{\alpha^2} = \frac{1}{\frac{\beta^2}{N} + \tau^2} \end{aligned}$$

Thus, W is an unbiased estimator of $1/(\beta^2/N+\tau^2)$, and can be used to replace $1/(\beta^2/N+\tau^2)$, in equation (10).

d. ESTIMATION OF β^2

It can be shown (*e.g.*, see Appendix A.7) that:

$$S_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_{ij} - Y_j)^2$$

is an unbiased estimator for β^2 based on the data from gene j, and that has a Chi-square distribution with (N-1) degrees of freedom. Since this is $\frac{N-1}{\beta^2}S_j^2$ the case for every j, a more accurate estimate for β^2 is obtained by pooling all available data, *i.e.*, by averaging the estimates for each j:

$$\begin{aligned}\widehat{\beta^2} &= \frac{1}{M}\sum_{j=1}^{M} S_j^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij} - Y_j)^2\right) \\ &= \frac{1}{M(N-1)}\sum_{j=1}^{M}\sum_{i=1}^{N}(X_{ij} - Y_j)^2.\end{aligned}$$

may be an unbiased estimator for β^2, because

$$\begin{aligned}\mathbf{E}(\widehat{\beta^2}) &= \mathbf{E}\left(\frac{1}{M}\sum_{j=1}^{M} S_j^2\right) \\ &= \frac{1}{M}\sum_{j=1}^{M}\mathbf{E}(S_j^2) = \frac{1}{M}\sum_{j=1}^{M}\beta^2 = \beta^2.\end{aligned}$$

Substituting the estimates (11) and (13) into equation (10), an explicit estimate for θ_j may be obtained:

$$
\begin{aligned}
&\widehat{\theta_j} \\
&= \left(1 - \frac{\widehat{1}}{\frac{\beta^2}{N} + \tau^2}\frac{\widehat{\beta^2}}{N}\right) Y_j \\
&= \left(1 - W \cdot \frac{\widehat{\beta^2}}{N}\right) Y_j \\
&= \left(1 - \left(\frac{M-2}{\sum_{k=1}^{M} {Y_k}^2}\right) \cdot \frac{1}{N} \cdot \frac{1}{M(N-1)} \sum_{k=1}^{M}\sum_{i=1}^{N} (X_{ik} - Y_k)^2\right) Y_j \\
&= \underbrace{\left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M}\sum_{i=1}^{N} (X_{ik} - Y_k)^2}{\sum_{k=1}^{M} {Y_k}^2}\right)}_{\gamma} Y_j \\
&= \gamma \overline{X}_{\cdot j}
\end{aligned}
\tag{14}
$$

Further, θ_j from equation (14) may be substituted into the correlation coefficient in equation (3) wherever $(X_j)_{offset}$ appears to obtain an explicit formula for $S(X_{\cdot j}, X_{\cdot k})$.

CLUSTERING

In an exemplary embodiment of the present invention, the genes may be clustered using the same hierarchical clustering algorithm as used by Eisen, except that G_{offset} is set equal to $\gamma\overline{G}$, where γ is a value between 0.0 and 1.0. The hierarchical clustering algorithm used by Eisen is based on the centroid-linkage method, which is referred to as "an average-linkage method" described in Sokal *et al.* ("Sokal"), *Univ. Kans. Sci. Bull.* 38, 1409-1438 (1958), the disclosure of which is incorporated herein by reference in its entirety. This method may compute a binary tree (dendrogram) that assembles all the genes at the leaves of the tree, with each internal node representing possible clusters at different levels. For any set of M genes, an upper-triangular similarity matrix may be computed by using a similarity metric of the type described in Eisen, which contains similarity scores for all pairs of genes. A node can be created joining the most similar pair of genes, and a gene

expression profile can be computed for the node by averaging observations for the joined genes. The similarity matrix may be updated with such new node replacing the two joined elements, and the process may be repeated (M -1) times until a single element remains. Because each internal node can be labeled by a value representing the similarity between its two children nodes (*i.e.*, the two elements that were combined to create the internal node), a set of clusters may be created by breaking the tree into subtrees (*e.g.*, by eliminating the internal nodes with labels below a certain predetermined threshold value). The clusters created in this manner can be used to compare the effects of choosing differing similarity measures.

III. ALGORITHM & IMPLEMENTATION

An exemplary implementation of a hierarchical clustering can proceed by selecting the most similar pair of elements (starting with genes at the bottom-most level) and combining them to create a new element. The "expression vector" for the new element can be the weighted average of the expression vectors of the two most similar elements that were combined. This exemplary structure of repeated pair-wise combinations may be represented in a binary tree, whose leaves can be the set of genes, and whose internal nodes can be the elements constructed from the two children nodes. The exemplary algorithm according to the present invention is described below in pseudocode.

HIERARCHICAL CLUSTERING PSEUDOCODE

Given $\left\{\left\{X_{ij}\right\}_{i=1}^{N}\right\}_{j=1}^{M}$

Switch:

Pearson: $\gamma = 1$;

Eisen: $\gamma = 0$;

Shrinkage: {

$$\text{Compute } W = (M-2) \Big/ \sum_{j=1}^{M} \overline{X}_{\cdot j}^{\,2}$$

$$\text{Compute } \widehat{\beta^2} = \sum_{j=1}^{M} \sum_{i=1}^{N} \left(X_{ij} - \overline{X}_{\cdot j}\right)^2 \Big/ \left(M(N-1)\right)$$

$$\gamma = 1 - W \cdot \widehat{\beta^2}/N$$

}

While (# clusters > 1) do

◊ Compute similarity table:

$$S(G_j, G_k) = \frac{\sum_i (G_{ij} - (G_j)_{offset})(G_{ik} - (G_k)_{offset})}{\sqrt{\sum_i (G_{ij} - (G_j)_{offset})^2 \cdot \sum_i (G_k)_{offset})^2}}, \quad (14)$$

where $(G_l)_{offset} = \gamma \overline{G}_l$.

◊ Find (j^*, k^*) :

$S(G_{j^*}, G_k^*) \geq S(G_j, G_k)$ $\forall$ clusters j, k

◊ Create new cluster $N_{j^*k^*}$

= weighted average of G_{j^*} and G_{k^*}.

◊ Take out clusters j^* and k^*.

IV. MATHEMATICAL SIMULATIONS AND EXAMPLES

a. *IN SILICO* EXPERIMENT

To compare the performance of these exemplary algorithms, it is possible to conduct an *in silico* experiment. In such an experiment, two genes X and Y can be created, and N (about 100) experiments can be simulated, as follows:

$$\begin{aligned} X_i &= \theta_X + \sigma_X(\alpha_i(X,Y) + \mathcal{N}(0,1)), \text{ and} \\ Y_i &= \theta_Y + \sigma_Y(\alpha_i(X,Y) + \mathcal{N}(0,1)), \end{aligned}$$

where α_i, chosen from a uniform distribution over a range $[L, H]$ ($U(L, H)$), can be a "bias term" introducing a correlation (or none if all α's are zero) between X and Y. $\theta_x \sim N(0,\tau^2)$ and $\theta_y \sim N(0,\tau^2)$, are the means of X and Y, respectively. Similarly, σx and σy are the standard deviations for X and Y, respectively.

With this model

$$
\begin{aligned}
S(X,Y) &= \frac{1}{N}\sum_{i=1}^{N}\frac{(X_i-\theta_X)}{\sigma_X}\frac{(Y_i-\theta_Y)}{\sigma_Y} \\
&\sim \frac{1}{N}\sum_{i=1}^{N}(\alpha_i+\mathcal{N}(0,1))(\alpha_i+\mathcal{N}(0,1)) \\
&\sim \frac{1}{N}\left[\left(\sum_{i=1}^{N}\alpha_i^2\right)+\chi^2_N+2\mathcal{N}(0,1)\sum_{i=1}^{N}\alpha_i\right]
\end{aligned}
$$

if the exact values of the mean and variance are used. The distribution of S is denoted by $F(\mu,\delta)$, where μ is the mean and δ is the standard deviation.

The model was implemented in Mathematica (*See* Wolfram ("Wolfram"), *The Mathematica Book.* Cambridge University Press, 4th Ed. (1999), the disclosure of which is incorporated herein by reference in its entirety). The following parameters were used in the simulation: $N = 100$, $\tau \in \{0.1, 10.0\}$ (representing very low or high variability among the genes), $\sigma_x = \sigma_Y = 10.0$, and $\alpha = 0$ representing no correlation between the genes or $\alpha \sim U(0, 1)$ representing some correlation between the genes. Once the parameters were fixed for a particular *in silico* experiment, the gene-expression vectors for X and Y were generated several thousand times, and for each pair of vectors $S_c(X, Y)$, $S_p(X, Y)$, $S_e(X, Y)$, and $S_s(X, Y)$ were estimated by four different algorithms and further examined to see how the estimators of S varied over these trials. These four different algorithms estimated S according to equations (1) and (2), as follows: *Clairvoyant* estimated S_c using the true values of θ_X, θ_Y, σ_X and σ_Y; *Pearson* estimated S_p using the unbiased estimators $\overline{X}$ and $\overline{Y}$ of σ_X, and σ_Y (for X_{offset} and Y_{offset}), respectively; *Eisen* estimated S_e using the value 0.0 as the estimator of both σ_X, and σ_Y; and *Shrinkage* estimated S_s using the shrunk biased estimators $\hat{\theta}_X$ and $\hat{\theta}_Y$ of θ_X and θ_Y, respectively. In the latter three, the standard deviation was estimated as in equation (2). The histograms corresponding to these in silica experiments can be found in Figure 4 (See Below). The information obtained from these conducted simulations, is as follows:

When X and Y are not correlated and the noise in the input is low (N = 100, τ = 0.1, and α = 0), Pearson performs about the same as Eisen, Shrinkage, and Clairvoyant ($S_c \sim F(-0.000297,0.0996)$, $S_p \sim F(-0.000269,0.0999)$, $S_e \sim F(-0.000254,0.0994)$, and $S_s \sim F(-0.000254,0.0994)$).

When X and Y are not correlated, but the noise in the input is high (N = 100, τ = 10.0, and α = 0), Pearson performs about as well as Shrinkage and Clairvoyant, but Eisen introduces a substantial number of false-positives ($S_c \sim F(-0.000971,0.0994)$, $Sp \sim F(-0.000939,0.100)$, $S_e \sim F(-0.00119, 0.354)$, and $S_s \sim F(-0.000939,0.100)$).

When X and Y are correlated and the noise in the input is low (N = 100, τ = 0.1, and $\alpha \sim U(0,1)$), Pearson performs substantially worse than Eisen, Shrinkage, and Clairvoyant, and Eisen, Shrinkage, and Clairvoyant perform about equally as well. Pearson introduces a substantial number of false-negatives ($S_c \sim F(0.331,0.132)$, $S_p \sim F(0.0755,0.0992)$, $S_e \sim F(0.248, 0.0915)$, and $S_s \sim F(0.245, 0.0915)$).

Finally, when X and Y are correlated and the noise in the input is high, the signal-to-noise ratio becomes extremely poor regardless of the algorithm employed ($S_c \sim F(0.333, 0.133)$, $S_p \sim F(0.0762,0.100)$, $S_e \sim F(0.117, 0.368)$, and $S_s \sim F(0.0762, 0.0999)$).

In summary, Pearson tends to introduce more false negatives and Eisen tends to introduce more false positives than Shrinkage. Exemplary Shrinkage procedures according to the present invention, on the other hand, can reduce these errors by combining the positive properties of both algorithms.

b. BIOLOGICAL EXAMPLE

Exemplary algorithms also were tested on a biological example. A biologically well-characterized system was selected, and the clusters of genes involved in the yeast cell cycle were analyzed. These clusters were computed using the hierarchical clustering algorithm with the underlying similarity measure chosen from the following three: Pearson, Eisen, or Shrinkage. As a reference, the computed clusters were compared to the ones implied by the common cell-cycle functions and

regulatory systems inferred from the roles of various transcriptional activators (See description associated with Figure 5 below).

The experimental analysis was based on the assumption that the groupings suggested by the ChIP (Chromatin ImmunoPrecipitation) analysis are correct and thus, provide a direct approach to compare various correlation coefficients. It is possible that the ChIP-based groupings themselves contain several false relations (both positives and negatives). Nevertheless, the trend of reduced false positives and false negatives using shrinkage analysis appears to be consistent with the mathematical simulation set forth above.

In Simon *et al.* ("Simon"), *Cell* 106, 697-708 (2001), the disclosure of which is incorporated herein by reference in its entirety, genome-wide location analysis is used to determine how the yeast cell cycle gene expression program is regulated by each of the nine known cell cycle transcriptional activators: Ace2, Fkhl, Fkh2, Mbpl, Mcml, Nddl, Swi4, Swi5, and Swi6. It was also determined that cell cycle transcriptional activators which function during one stage of the cell cycle regulate transcriptional activators that function during the next stage. According to an exemplary embodiment of the present invention, these serial regulation transcriptional activators, together with various functional properties, can be used to partition some selected cell cycle genes into nine clusters, each one characterized by a group of transcriptional activators working together and their functions (see Table 1). For example, Group 1 may characterized by the activators Swi4 and Swi6 and the function of budding; Group 2 may be characterized by the activators Swi6 and Mbpl and the function involving DNA replication and repair at the juncture of G1 and S phases, *etc.*

The hypothesis in this exemplary embodiment of the present invention can be summarized as follows: genes expressed during the same cell cycle stage (and regulated by the same transcriptional activators) can be in the same cluster. Provided below are exemplary deviations from this hypothesis that are observed in the raw data.

<u>Possible False Positives</u>:

Bud9 (Group 1: Budding) and {Ctsl, Egt2} (Group 7: Cytokinesis) can be placed in the same cluster by all three metrics: P49 = S82 ≃ E47; however, the Eisen metric also places Exgl (Group 1) and Cdc6 (Group 8: Pre-replication complex formation) in the same cluster.

Mcm2 (Group 2: DNA replication and repair) and Mcm3 (Group 8) can be placed in the same cluster by all three metrics: P10 = S20 ≃ E73; however, the Eisen metric places several more genes from different groups in the same cluster: {Rnr1, Rad27, Cdc21, Dun1, Cdc45} (Group 2), Hta3 (Group 3: Chromatin), and Mcm6 (Group 8) are also placed in cluster E73.

Table 1: Genes in our data set, grouped by transcriptional activators and cell-cycle functions.

	Activators	Genes	Functions
1	Swi4, Swi6	Cln1, Cln2, Gic1, Gic2, Msb2, Rsr1, Bud9, Mnn1, Och1, Exg1, Kre6, Cwp1	Budding
2	Swi6, Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2	DNA replication and repair
3	Swi4, Swi6	Htb1, Htb2, Hta1, Hta2, Hta3, Hho1	Chromatin
4	Fkh1	Hhf1, Hht1, Tel2, Arp7	Chromatin
5	Fkh1	Tem1	Mitosis Control
6	Ndd1, Fkh2, Mcm1	Clb2, Ace2, Swi5, Cdc20	Mitosis Control
7	Ace2, Swi5	Cts1, Egt2	Cytokinesis
8	Mcm1	Mcm3, Mcm6, Cdc6, Cdc46	Pre-replication complex formation
9	Mcm1	Ste2, Far1	Mating

Possible False Negatives:

Group 1: Budding (Table 1) may be split into four clusters by the Eisen metric: {Cln1, Cln2, Gic2, Rsr1, Mnn1} ∈ Cluster *a* (E39), Gic2 ∈ Cluster *b* (E62), {Bud9, Exg1} ∈ Cluster *c* (E47), and {Kre6, Cwp1} ∈ Cluster d (E66); and into six clusters by both the Shrinkage and Pearson metrics: {Cln1, Cln2, Gic2, Rsr1, Mnn1}

∈ Cluster *a* (S3=P66), {Gic1, Kre6} ∈ Cluster *b* (S39=PI7), Msb2 ∈ Cluster *c* (S24=P71), Bud9 ∈ Cluster *d* (S82=P49), Exg1 ∈ Cluster *e* (S48=P78), and Cwp1 ∈ Cluster *f* (S8=P4).

Table 1 contains those genes from Figure 5 that were present in an evaluated data set. The following tables contain these genes grouped into clusters by an exemplary hierarchical clustering algorithm according to the present invention using the three metrics (Eisen in Table 2, Pearson in Table 3, and Shrinkage in Table 4) threshold at a correlation coefficient value of 0.60. The choice of the threshold parameter is discussed further below. Genes that have not been grouped with any others at a similarity of 0.60 or higher are not included in the tables. In the subsequent analysis they can be treated as *singleton* clusters.

Table 2: Eisen Clusters

E39	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
E62	Swi4/Swi6	Gic1
E47	Swi4/Swi6 Ace2/Swi5 Mcm1	Bud9, Exg1 Cts1, Egt2 Cdc6
E66	Swi4/Swi6	Kre6, Cwp1
E71	Swi6/Mbp1 Fkh1 Ndd1/Fkh2/Mcm1 Mcm1	Clb5, Clb6, Rad51 Tel2 Cdc20 Cdc46
E73	Swi6/Mbp1 Swi4/Swi6 Mcm1	Rnr1, Rad27, Cdc21, Dun1, Cdc45, Mcm2 Hta3 Mcm3, Mcm6
E63	Swi4/Swi6 Fkh1	Htb1, Htb2, Hta1, Hta2, Hhc1 Hhf1, Hhf2
E32	Fkh1	Arp7
E38	Fkh1 Ndd1/Fkh2/Mcm1	Tem1 Clb2, Ace2, Swi5
E51	Mcm1	Ste2, Far1

Table 3: Pearson Clusters

P66	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
P17	Swi4/Swi6	Gic1, Kre6
P71	Swi4/Swi6	Msb2
P49	Swi4/Swi6 Ace2/Swi5	Bud9 Cts1, Egt2
P78	Swi4/Swi6	Exg1
P4	Swi4/Swi6	Cwp1
P12	Swi5/Mbp1 Swi4/Swi6 Fkh1 Ndd1/Fkh2/Mcm1 Mcm1	Clb5, Clb6, Rnr1, Cdc21, Dun1, Rad51, Cdc45 Hta3 Tel2 Cdc20 Mcm6, Cdc46
P10	Swi6/Mbp1 Mcm1	Mcm2 Mcm3
P54	Swi4/Swi6 Fkh1	Htb1, Htb2, Hta1, Hta2, Hho1 Hhf1, Hht1
P37	Fkh1	Arp7
P16	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
P50	Mcm1	Ste2, Far1

Table 4: Shrinkage Clusters

S3	Swi4/Swi6	Cln1, Cln2, Gic2, Rsr1, Mnn1
S39	Swi4/Swi6	Gic1, Kre6
S21	Swi4/Swi6	Msb2
S32	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
S18	Swi4/Swi6	Exg1
S8	Swi4/Swi6	Cwp1
S14	Swi6/Mbp1	Clb5, Clb6, Rnr1, Cdc21, Dun1, Rad51, Cdc45
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
S20	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S4	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S13	Swi4/Swi6	Hta3
S63	Fkh1	Arp7
S22	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S33	Mcm1	Ste2, Far1

The value $\gamma = 0.89$ estimated from the raw yeast data appears to be greater than a γ value based equation [1]. Moreover, the value $\gamma = 0$ performed better than $\gamma = 1$. Such value also appears not to have yielded as great an improvement in the yeast data clusters as the simulations indicated. This exemplary result indicates that the true value of γ may be closer to 0. Upon a closer examination of the data, it can be observed that it may be possible that the data in its raw "pre-normalized" form is inconsistent with the assumptions used in deriving γ:

1. The gene vectors are not range-normalized, so $\beta_j^2 \neq \beta^2$ for every j; and

2. The N experiments are not necessarily independent.

CORRECTIONS

The first observation may be compensated for by normalizing all gene vectors with respect to range (dividing each entry in gene X by $(X_{max} - X_{min})$), recomputing the estimated, value, and repeating the clustering process. As normalized gene expression data yielded the estimate $\gamma \cong 0.91$ appears to be too high

a value, an extensive computational experiment was conducted to determine the best empirical γ value by also clustering with the shrinkage factors of 0.2, 0.4, 0.6, and 0.8. The clusters taken at the correlation factor cut-off of 0.60, as above, are presented in Tables 5, 6, 7, 8, 9, 10 and 11.

Table 5: RN Data, $\gamma = 0.0$ (Eisen Clusters)

E8	Swi4/Swi6	Cln1, Msb2, Mnn1
E71	Swi4/Swi6	Cln2, Rsr1
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
E14	Swi4/Swi6	Gic1
E17	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
	Mcm1	Ste2, Far1
E16	Swi4/Swi6	Exg1
E59	Swi4/Swi6	Kre6
E18	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
E86	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hhb1
	Fkh1	Hhf1, Hht1
E10	Fkh1	Arp7
E19	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
E11	Mcm1	Cdc6

Table 6: Range-normalized data, $\gamma = 0.2$

$S_{0.2}59$	Swi4/Swi6	Cln1, Cic2, Rsr1, Mnn1
$S_{0.2}26$	Swi4/Swi6 Swi6/Mbp1	Cln2 Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cde45
$S_{0.2}23$	Swi4/Swi6	Gic1
$S_{0.2}53$	Swi4/Swi6 Ace2/Swi5	Bud9 Cts1, Egt2
$S_{0.2}57$	Swi4/Swi6 Fkh1	Exg1 Arp7
$S_{0.2}61$	Swi4/Swi6	Kre6
$S_{0.2}18$	Swi6/Mbp1 Swi4/Swi6 Fkh1 Ndd1/Fkh2/Mcm1 Mcm1	Clb5 Hta3 Tel2 Cdc20 Mcm6, Cde46
$S_{0.2}28$	Swi6/Mbp1 Mcm1	Mcm2 Mcm3
$S_{0.2}25$	Swi4/Swi6 Fkh1	Htb1, Htb2, Hta1, Htn2, Hho1 Hhf1, Hht1
$S_{0.2}29$	Fkh1 Ndd1/Fkh2/Mcm1	Tem1 Clb2, Ace2, Swi5
$S_{0.2}4$	Mcm1	Ste2
$S_{0.2}55$	Mcm1	Far1

Table 7: Range-normalized data, $\gamma = 0.4$

$S_{0.4}64$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.4}13$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.4}44$	Swi4/Swi6	Gic1, Krs6
$S_{0.4}27$	Swi4/Swi6	Msb2
$S_{0.4}46$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.4}73$	Swi4/Swi6	Exg1
$S_{0.4}2$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.4}48$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.4}26$	Fkh1	Arp7
$S_{0.4}25$	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.4}16$	Mcm1	Cdc6
$S_{0.4}47$	Mcm1	Ste2
$S_{0.4}58$	Mcm1	Far1

Table 8: Range-normalized data, $\gamma = 0.6$

$S_{0.6}34$	Swi4/Swi6	Cln1, Gic2, Rsr1, Mnn1
$S_{0.6}77$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.6}35$	Swi4/Swi6	Gic1, Kre6
$S_{0.6}47$	Swi4/Swi6	Msb2
$S_{0.6}62$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.6}20$	Swi4/Swi6	Exg1
$S_{0.6}73$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.6}91$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hbo1
	Fkh1	Hhf1, Hhf2
$S_{0.6}48$	Fkh1	Arp7
$S_{0.6}37$	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.6}64$	Mcm1	Ste2
$S_{0.6}63$	Mcm1	Far1

Table 9: Range-normalized data, $\gamma = 0.8$

$S_{0.8}51$	Swi4/Swi6	Cln1, Gic2, Rar1, Mnn1
$S_{0.8}7$	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
$S_{0.8}64$	Swi4/Swi6	Gic1, Kre6
$S_{0.8}90$	Swi4/Swi6	Msb2
$S_{0.8}31$	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
$S_{0.8}43$	Swi4/Swi6	Exg1
$S_{0.8}65$	Swi4/Swi6	Cwp1
$S_{0.8}13$	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
$S_{0.8}17$	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
$S_{0.8}76$	Fkh1	Arp7
$S_{0.8}74$	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
$S_{0.8}33$	Mcm1	Sbe2
$S_{0.8}32$	Mcm1	Far1

Table 10: RN Data, $\gamma = 0.91$ (Shrinkage Clusters)

S49	Swi4/Swi6	Cln1, Gic2, Rsr1, Mun1
S73	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
S45	Swi4/Swi6	Gic1, Kre6
S15	Swi4/Swi6	Msb2
S90	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
S56	Swi4/Swi6	Exg1
S46	Swi4/Swi6	Cwp1
S71	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S61	Swi4/Swi6	Hth1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hhf1
S37	Fkh1	Arp7
S7	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S91	Mcm1	Ste2
S92	Mcm1	Far1

Table 11: RN Data, $\gamma = 1.0$ (Pearson Clusters)

P10	Swi4/Swi6	Clu1, Gic2, Rsr1, Mun1
P68	Swi4/Swi6	Cln2
	Swi6/Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
	Swi4/Swi6	Hta3
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Mcm6, Cdc46
P1	Swi4/Swi6	Gic1, Kre6
P39	Swi4/Swi6	Msb2
P66	Swi4/Swi6	Bud9
	Ace2/Swi5	Cts1, Egt2
P20	Swi4/Swi6	Exg1
P2	Swi4/Swi6	Cwp1
P72	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
P53	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hbs1
	Fkh1	Hhf1, Hhf1
P12	Fkh1	Arp7
P46	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
P61	Mcm1	Ste2
P65	Mcm1	Far1

To compare the resulting sets of clusters, the following notation may be introduced. Each cluster set may be written, as follows:

$$\left\{x \rightarrow \{\{y_1, z_1\}, \{y_2, z_2\}, \ldots, \{y_{n_x}, z_{n_x}\}\}\right\}_{x=1}^{\#\text{ of groups}}$$

where x denotes the group number (as described in Table 1), n_x is the number of clusters group x appears in, and for each cluster $j \in \{1, \ldots, n_x\}$, where are y_j genes from group x and z_j genes from other groups in Table 1. A value of "*" for z_j denotes that cluster j contains additional genes, although none of them are cell cycle genes; in subsequent computations, this value may be treated as 0.

This notation naturally lends itself to a scoring function for measuring the number of false positives, number of false negatives, and total error score, which aids in the comparison of cluster sets.

$$\mathrm{FP}(\gamma) = \frac{1}{2}\sum_{x}\sum_{j=1}^{n_x} y_j \cdot z_j \qquad (15)$$

$$\mathrm{FN}(\gamma) = \sum_{x}\sum_{1 \le j(k \le n_x} y_j \cdot y_k \qquad (16)$$

$$\mathrm{Error\_score}(\gamma) = \mathrm{FP}(\gamma) + FN(\gamma) \qquad (17)$$

$\gamma = 0.0(E) \Longrightarrow$
{1 → {{3, *}, {2, 13}, {1, *}, {1, *},
{1, *}, {1, 4}, {1, 0}, {1, 0}, {1, 0}},
2 → {{8, 7}, {1, 1}},
3 → {{5, 2}, {1, 14}},
4 → {{2, 5}, {1, 14}, {1, *}},
5 → {{1, 3}},
6 → {{8, 1}, {1, 14}},
7 → {{2, 3}},
8 → {{2, 13}, {1, 1}, {1, 0}},
9 → {{2, 3}}
}
Error_score(0.0) = 97 + 88 = 185

$\gamma = 0.2 \Longrightarrow$

$\{1 \rightarrow \{\{4,*\},\{1,7\},\{1,*\},\{1,*\},$
$\{1,1\},\{1,2\},\{1,0\},\{1,0\},\{1,0\}\},$
$2 \rightarrow \{\{7,1\},\{1,5\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,5\}\},$
$4 \rightarrow \{\{2,5\},\{1,5\},\{1,1\}\},$
$5 \rightarrow \{\{1,3\}\},$
$6 \rightarrow \{\{3,1\},\{1,5\}\},$
$7 \rightarrow \{\{2,1\}\},$
$8 \rightarrow \{\{2,4\},\{1,1\},\{1,0\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(0.2) = 38 + 94 = 132

In such notation, the cluster sets with their error scores can be listed as follows:

$\gamma = 0.4 \Longrightarrow$

$\{1 \rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},$
$\{2,*\},\{1,2\},\{1,0\},\{1,0\}\},$
$2 \rightarrow \{\{8,6\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,13\}\},$
$4 \rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},$
$5 \rightarrow \{\{1,3\}\},$
$6 \rightarrow \{\{3,1\},\{1,13\}\},$
$7 \rightarrow \{\{2,1\}\},$
$8 \rightarrow \{\{2,12\},\{1,*\},\{1,1\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(0.4) = 78 + 86 = 164

$\gamma = 0.6 \Longrightarrow$

$\{1 \rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},$
$\{2,*\},\{1,2\},\{1,0\},\{1,0\}\},$
$2 \rightarrow \{\{8,6\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,13\}\},$
$4 \rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},$
$5 \rightarrow \{\{1,0\}\},$
$6 \rightarrow \{\{3,*\},\{1,13\}\},$
$7 \rightarrow \{\{2,1\}\},$
$8 \rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(0.6) = 75 + 86 = 161

Error_score(0.6) = 75 + 86 = 161.

$\gamma = 0.91(S) \Longrightarrow$

$\{1 \rightarrow \{\{4,*\},\{1,13\}\{1,*\},\{1,*\},$
$\{1,*\},\{2,*\},\{1,3\},\{1,0\}\},$
$2 \rightarrow \{\{8,6\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,13\}\},$
$4 \rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},$
$5 \rightarrow \{\{1,0\}\},$
$6 \rightarrow \{\{3,*\},\{1,13\}\},$
$7 \rightarrow \{\{2,1\}\},$
$8 \rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

$\gamma = 0.8 \Longrightarrow$

$\{1 \rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},$
$\{1,*\},\{2,*\},\{1,2\},\{1,0\}\},$
$2 \rightarrow \{\{8,6\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,13\}\},$
$4 \rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},$
$5 \rightarrow \{\{1,0\}\},$
$6 \rightarrow \{\{3,*\},\{1,13\}\},$
$7 \rightarrow \{\{2,1\}\},$
$8 \rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(0.8) = 75 + 86 = 161

Error_score(0.91) = 75 + 86 = 161.

$\gamma = 1.0 (P) \Longrightarrow$

$\{1 \rightarrow \{\{4,*\},\{1,13\},\{1,*\},\{1,*\},$
$\{1,*\},\{2,*\},\{1,2\},\{1,0\}\},$
$2 \rightarrow \{\{8,6\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,13\}\},$
$4 \rightarrow \{\{2,5\},\{1,13\},\{1,*\}\},$
$5 \rightarrow \{\{1,0\}\},$
$6 \rightarrow \{\{3,*\},\{1,13\}\},$
$7 \rightarrow \{\{2,1\}\},$
$8 \rightarrow \{\{2,12\},\{1,1\},\{1,0\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(1.0) = 75 + 86 = 161

In this notion, γ values of 0.8, 0.91, and 1.0 provide substantially identical cluster groupings, and the likely best error score may be attained at $\gamma = 0.2$.

To improve the estimated value of γ, the statistical dependence among the experiments may be compensated for by reducing the effective number of experiments by subsampling from the set of all (possibly correlated) experiments. The candidates can be chosen via clustering all the experiments, *i.e.*, columns of the data matrix, and then selecting one representative experiment from each cluster of experiments. The subsampled data may then be clustered, once again using the cut-

off correlation value of 0.60. The exemplary resulting cluster sets under the Eisen, Shrinkage, and Pearson metrics are given in Tables 12, 13, and 14, respectively.

Table 12: RN Subsampled Data, $\gamma = 0.0$ (Eisen)

E58	Swi4/Swi6	Cln1, Och1
E68	Swi4/Swi6	Cln2, Msb2, Rsr1, Bud9, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2
	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hhn1
	Fkh1	Hhf1, Hht1, Arp7
	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
	Ace2/Swi5	Egt2
	Mcm1	Mcm3, Mcm6, Cdc6
E29	Swi4/Swi6	Cic1
E61	Swi4/Swi6	Cic2
E33	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Swi4/Swi6	Hta3
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
E73	Fkh1	Tel2
E23	Ace2/Swi5	Cts1
E43	Mcm1	Ste2
E66	Mcm1	Fur1

Table 13: RN Subsampled Data, $\gamma = 0.66$ (Shrinkage)

S49	Swi4/Swi6	Cln1, Bud9, Och1
	Ace2/Swi5	Egt2
	Mcm1	Cdc6
S6	Swi4/Swi6	Cln2, Gic2, Msb2, Rsr1, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
S32	Swi4/Swi6	Gic1
S65	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Fkh1	Tel2
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
S15	Swi6/Mbp1	Mcm2
	Mcm1	Mcm3
S11	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1
S60	Swi4/Swi6	Hta3
S30	Fkh1	Arp7
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
S62	Fkh1	Tem1
S53	Ace2/Swi5	Cts1
S14	Mcm1	Mcm6
S35	Mcm1	Ste2
S36	Mcm1	Far1

Table 14: RN Subsampled Data, $\gamma = 1.0$ (Pearson)

P1	Swi4/Swi6	Cln1, Och1
P15	Swi4/Swi6 Swi6/Mbp1 Mcm1	Cln2, Rsr1, Mnn1 Cdc21, Dun1, Rad51, Cdc45, Mcm2 Mcm3
P29	Swi4/Swi6	Gin1
P2	Swi4/Swi6	Gin2
P3	Swi4/Swi6 Swi6/Mbp1	Msb2, Exg1 Rnr1
P51	Swi4/Swi6 Ndd1/Fkh2/Mcm1 Ace2/Swi5 Mcm1	Bud9 Clb2, Ace2, Swi5 Hgt2 Cdc6
P11	Swi4/Swi6	Kre6
P62	Swi4/Swi6 Swi6/Mbp1 Swi4/Swi6 Ndd1/Fkh2/Mcm1 Mcm1	Cwp1 Clb5, Clb6 Hta3 Cdc20 Cdc46
P49	Swi6/Mbp1 Swi4/Swi6 Fkh1	Rad27 Htb1, Htb2, Hta1, Hta2, Hhb1 Hhf1, Hhf1
P10	Fkh1 Mcm1	Tel2 Mcm6
P23	Fkh1	Arp7
P50	Fkh1	Tem1
P69	Ace2/Swi5	Cts1
P42	Mcm1	Ste2
P13	Mcm1	Far1

The subsampled data may yield the lower estimated value ≈ 0.66. In the exemplary set notation, the resulting clusters with the corresponding error scores can be written as follows:

$$\gamma = 0.0(E) \Longrightarrow$$

$$\begin{aligned}
\{1 &\rightarrow \{\{6,23\},\{2,*\},\{2,5\},\{1,*\},\{1,*\}\},\\
2 &\rightarrow \{\{7,22\},\{2,5\}\},\\
3 &\rightarrow \{\{5,24\},\{1,6\}\},\\
4 &\rightarrow \{\{3,26\},\{1,*\}\},\\
5 &\rightarrow \{\{1,28\}\},\\
6 &\rightarrow \{\{3,26\},\{1,6\}\},\\
7 &\rightarrow \{\{1,*\},\{1,28\}\},\\
8 &\rightarrow \{\{3,26\},\{1,6\}\},\\
9 &\rightarrow \{\{1,*\},\{1,*\}\}\\
\}&
\end{aligned}$$

$$\text{Error\_score}(0.0) = 370 + 79 = 449$$

$\gamma = 0.66(S) \Longrightarrow$

$\{1 \rightarrow \{\{6,6\},\{3,2\},\{2,5\},\{1,*\}\}, \quad \{1,*\}\},$
$2 \rightarrow \{\{6,6\},\{2,5\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,*\}\},$
$4 \rightarrow \{\{2,5\},\{1,3\},\{1,6\}\},$
$5 \rightarrow \{\{1,*\}\},$
$6 \rightarrow \{\{3,1\},\{1,6\}\},$
$7 \rightarrow \{\{1,*\},\{1,4\}\},$
$8 \rightarrow \{\{1,*\},\{1,1\},\{1,4\},\{1,6\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(0.66) = 76 + 88 = 164

$\gamma = 1.0(P) \Longrightarrow$

$\{1 \rightarrow \{\{3,6\},\{2,*\},\{2,1\},\{1,*\},$
$\{1,*\},\{1,*\},\{1,5\},\{1,5\}\},$
$2 \rightarrow \{\{5,4\},\{2,4\},\{1,2\},\{1,7\}\},$
$3 \rightarrow \{\{5,3\},\{1,5\}\},$
$4 \rightarrow \{\{2,6\},\{1,*\},\{1,1\}\},$
$5 \rightarrow \{\{1,*\}\},$
$6 \rightarrow \{\{3,3\},\{1,5\}\},$
$7 \rightarrow \{\{1,*\},\{1,5\}\},$
$8 \rightarrow \{\{1,1\},\{1,5\},\{1,5\},\{1,8\}\},$
$9 \rightarrow \{\{1,*\},\{1,*\}\}$
$\}$

Error_score(1.0) = 69 + 107 = 176

From the tables for the range-normalized, subsampled yeast data, as well as by comparing the error scores, it appears that for the same clustering algorithm and threshold value, Pearson introduces more false negatives and Eisen introduces more false positives than Shrinkage. The exemplary Shrinkage procedure according to the present invention may reduce these errors by combining the positive properties of both algorithms. This observation is consistent with the mathematical analysis and simulation described above.

GENERAL DISCUSSION

Microarray-based genomic analysis and other similar high-throughput methods have begun to occupy an increasingly important role in biology, as they have helped to create a visual image of the state-space trajectories at the core of the cellular processes. Nevertheless, as described above, a small error in the estimation of a parameter (*e.g.*, the shrinkage parameter) may have a significant effect on the overall conclusion. Errors in the estimators can manifest themselves by missing certain biological relations between two genes (false negatives) or by proposing phantom relations between two otherwise unrelated genes (false positives).

A global illustration of these interactions can be seen in an exemplary Receiver Operator Characteristic ("ROC") graph (shown in Figure 6) with each curve parameterized by the cut-off threshold in the range of [-1,1]. The ROC curve (*see, e.g.*, Egan, J.P., *Signal Detection Theory and ROC analysis*, Academic Press, New York. (1975), the entire disclosure of which is incorporated herein by reference in its entirety) for a given metric preferably plots sensitivity against (1-specificity), where:

Sensitivity = fraction of positives detected by a metric

$$= \frac{TP(\gamma)}{TP(\gamma) + FN(\gamma)},$$

Specificity = fraction of negatives detected by a metric

$$= \frac{TN(\gamma)}{TN(\gamma) + FP(\gamma)},$$

and TP(γ), FN(γ), FP(γ) and TN(γ) denote the number of True Positives, False Negatives, False Positives, and True Negatives, respectively, arising from a metric associated with a given γ. (Recall that γ is 0.0 for Eisen, 1.0 for Pearson, and may be computed according to equation (14) for Shrinkage, which yields about 0.66 on this data set.) For each pair of genes, $\{j,k\}$, we can define these events using our hypothesis as a measure of truth:

TP: $\{j, k\}$ can be in same group (see Table 1) and $\{j, k\}$ can be placed in same cluster;

FP: $\{j, k\}$ can be in different groups, but $\{j, k\}$ can be placed in same cluster;

TN: $\{j, k\}$ can be in different groups and $\{j, k\}$ can be placed in different clusters; and

FN: $\{j, k\}$ can be in same group, but $\{j, k\}$ can be placed in different clusters.

FP(γ) and FN(γ) were already defined in equations (15) and (16), respectively, and we define

$$TP(\gamma) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2} \tag{18}$$

and

$$TN(\gamma) = Total - (TP(\gamma) + FN(\gamma) + FP(\gamma)) \tag{19}$$

where Total = $\binom{44}{2} = 946$ is the total # of gene pairs $\{j, k\}$ in Table 1.

The ROC figure suggests the best threshold to use for each metric, and can also be used to select the best metric to use for a particular sensitivity.

The dependence of the error scores on the threshold can be more clearly seen from an exemplary graph of Figure 7, which shows that a threshold value of about 0.60 is a reasonable representative value.

B. FINANCIAL EXAMPLE

The algorithms of the present invention may also be applied to financial markets. For example, the algorithm may be applied to determine the behavior of individual stocks or groups of stocks offered for sale on one or more publicly-traded stock markets relative to other individual stocks, groups of stocks, stock market indices calculated from the values of one or more individual stocks, *e.g.*, the Dow Jones 500, or stock markets as a whole. Thus, an individual considering investment in a given stock or groups of stocks in order to achieve a return on their investment greater than that provided by another stock, another group of stocks, a stock index or the market as a whole, could employ the algorithm of the present invention to determine whether the sales price of the given stock or group of stocks under consideration moves in a correlated way to the movement of any other stock, groups of stocks, stock indices or stock markets as a whole. If there is a strong association between the movement of the price of a given stock or groups of stocks and another stock, another group of stocks, a stock index or the market as a whole, the

prospective investor may not wish to assume the potentially greater risk associated with investing in a single stock when its likelihood to increase in value may be limited by the movement of the market as a whole, which is usually a less risky investment. Alternatively, an investor who knows or believes that a given stock has in the past outperformed other stocks, a stock market index, or the market as a whole, could employ the algorithm of the present invention to identify other promising stocks that are likely to behave similarly as future candidates for investment. Those skilled in the art of investment will recognize that the present invention may be applied in numerous systems, methods, and software arrangements for identifying candidate investments, not only in stock markets, but also in other markets including but not limited to the bond market, futures markets, commodities markets, etc., and the present invention is in no way limited to the exemplary applications and embodiments described herein.

The foregoing merely illustrates the principles of the present invention. Various modifications and alterations to the described embodiments will be apparent to those skilled in the art in view of the teachings herein. It will thus be appreciated that those skilled in the art will be able to devise numerous systems, methods, and software arrangements for determining associations between one or more elements contained within two or more datasets that, although not explicitly shown or described herein, embody the principles of the invention and are thus within the spirit and scope of the invention. Indeed, the present invention is in no way limited to the exemplary applications and embodiments thereof described above.

APPENDIX

APPENDIX A.1 - RECEIVER OPERATOR CHARACTERISTIC CURVES

Definitions

If two genes are in the same group, they may "belong in same cluster", and if they are in different groups, they may "belong in different clusters." Receiver Operator Characteristic (ROC) curves, a graphical representation of the number of true positives versus the number of false positives for a binary classification system as the discrimination threshold is varied, are generated for each metric used (*i.e.*, one for Eisen, one for Pearson, and one for Shrinkage).

Event: grouping of (cell cycle) genes into clusters;

Threshold: cut-off similarity value at which the hierarchy tree is cut into clusters.

The exemplary cell-cycle gene table can consist of 44 genes, which gives us C(44,2) = 946 gene pairs. For each (unordered) gene pair $\{j, k\}$, define the following events:

TP: $\{j, k\}$ can be in same group and $\{j, k\}$ can be placed in same cluster;

FP: $\{j, k\}$ can be in different groups, but $\{j, k\}$ can be placed in same cluster;

TN: $\{j,k\}$ can be in different groups and $\{j,k\}$ can be placed in different clusters; and

FN: $\{j, k\}$ can be in same group, but $\{j, k\}$ can be placed in different clusters.

Thus,

$$TP(\gamma) = \sum_{\{j,k\}} TP(\{j,k\})$$

$$FP(\gamma) = \sum_{\{j,k\}} FP(\{j,k\})$$

$$TN(\gamma) = \sum_{\{j,k\}} TN(\{j,k\})$$

$$FN(\gamma) = \sum_{\{j,k\}} FN(\{j,k\})$$

where the sums are taken over all 946 unordered pairs of genes.

Two other quantities involved in ROC curve generation can be

Sensitivity = fraction of positives detected by a metric

$$= \frac{TP(\gamma)}{TP(\gamma) + FN(\gamma)}.$$

Specificity = fraction of negatives detected by a metric

$$= \frac{TN(\gamma)}{TN(\gamma) + FP(\gamma)}.$$

The ROC curve plots sensitivity, on the *y*-axis, as a function of (1- specificity), on the *x*-axis, with each point on the plot corresponding to a different cut-off value. A different curve was created for each of the three metrics.

The following sections describe how the quantities TP(γ), FN(γ), FP(γ), and TN(γ) can be computed using an exemplary set notation for clusters, with a relationship of:

$$\left\{x \rightarrow \{\{y_1, z_1\}, \{y_2, z_2\}, \ldots, \{y_{n_x}, z_{n_x}\}\}\right\}_{x=1}^{\#\text{ of groups}}$$

Computations

A. TP

$$TP(\gamma) = \sum_{\{j,k\}} TP(\{j,k\}) =$$

gene pairs that were placed in same cluster

and belong in same group.

For each group *x* given in set notation as

$$x \rightarrow \{\{y_1, z_1\}, \ldots, \{y_{n_x}, z_{n_x}\}\},$$

pairs from each y_j should be counted, *i.e.*,

$$TP(x) = \binom{y_1}{2} + \ldots + \binom{y_{nx}}{2} = \sum_{j=1}^{n_x} \binom{y_j}{2}$$

Obtaining a total over all groups yields

$$TP(\gamma) = \sum_{x=1}^{\#groups} TP(x) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2}$$

B. FN

$$FN(\gamma) = \sum_{\{j,k\}} FN(\{j,k\}) =$$

gene pairs that belong in same group

but were placed into different clusters.

$$FN(x) = \begin{cases} \sum_{j=1}^{n_x} \sum_{k=j+1}^{n_x} y_j \cdot y_k & \text{if } n_x \geq 2, \text{ or} \\ 0, & \text{if } n_x = 1. \end{cases}$$

Every pair that was separated could be counted

However, when $n_x = 1$, there is no pair $\{j, k\}$ that satisfies the triple inequality $1 \leq j < k \leq n_x$, and hence, it is not necessary to treat such pair as a special case.

$$\therefore FN(\gamma) = \sum_{x=1}^{\#groups} FN(x) = \sum_{x} \sum_{1 \leq j < k \leq n_x} y_j \cdot y_k$$

C. FP

$$FP(\gamma) = \sum_{\{j,k\}} FP(\{j,k\}) =$$

gene pairs that belong in different groups but got placed in the same cluster.

The expression

$$\sum_{x} \sum_{j=1}^{n_x} y_j \cdot z_j$$

may count every false-positive pair *{j, k}* twice: first, when looking at *j*'s group, and again, when looking at k's group.

$$\therefore FP(\gamma) = \frac{1}{2} \sum_{x} \sum_{j=1}^{n_x} y_j \cdot z_j$$

D. TN

$$TN(\gamma) = \sum_{\{j,k\}} TN(\{j,k\}) =$$

gene pairs that belong in different groups and got placed in different clusters. Instead of counting true-negatives from our notation, the fact that the other three scores are known may be used, and the total thereof can also be utilized.

Complementarily. Given a gene pair $\{j,k\}$, only one of the events $\{TP(\{j,k\}), FN(\{j,k\}), FP(\{j,k\}), TN(\{j,k\})\}$ may be true. This implies

$$\sum_{\{j,k\}} TP(\{j,k\}) + \sum_{\{j,k\}} FN(\{j,k\}) +$$

$$+ \sum_{\{j,k\}} FP(\{j,k\}) + \sum_{\{j,k\}} TN(\{j,k\}) =$$

$$= TP(\gamma) + FN(\gamma) + FP(\gamma) + TN(\gamma) =$$

$$= \binom{44}{2} = \frac{44 \cdot 43}{2} = 946 = \text{Total}$$

$$\therefore TN(\gamma) = Total - (TP(\gamma) + FN(\gamma) + FP(\gamma))$$

Plotting ROC curves

For each cut-off value θ, TP(γ), FN(γ), FP(γ), and TN(γ) are computed as described above, with $\gamma \in \{0.0, 0.66, 1.0\}$ corresponding to Eisen, Shrinkage, and Pearson, respectively. Then, the sensitivity and specificity may be computed from equations (20) and (21), and sensitivity vs. (1-specificity) can be plotted, as shown in Figure 6.

The effect of the cut-off threshold θ on the FN and FP scores individually also can be examined, using an exemplary graph shown in Figure 7.

A 3-dimensional graph of (1-specificity) on the *x*–axis, sensitivity on the *y*–axis, and threshold on the *z*–axis offers a view shown in Figure 8.

A.2 Computing the marginal pdf for X_j

$$\begin{aligned} f(X_j) &= \mathbf{E}_{\theta_j} f(X_j|\theta_j) = \int_{-\infty}^{\infty} f(X_j|\theta)\pi(\theta)d\theta \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_j-\theta)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{\theta^2}{2\tau^2}} d\theta \\ &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{(X_j-\theta)^2}{\sigma^2}+\frac{\theta^2}{\tau^2}\right)} d\theta \end{aligned} \tag{22}$$

First, rewrite the exponent as a complete square:

$$\begin{aligned} \frac{(X_j-\theta)^2}{\sigma^2}+\frac{\theta^2}{\tau^2} &= \frac{1}{\sigma^2\tau^2}\left[\tau^2(X_j-\theta)^2+\sigma^2\theta^2\right] \\ &= \frac{1}{\sigma^2\tau^2}\left[\tau^2 X_j{}^2-2\tau^2 X_j\theta+\tau^2\theta^2+\sigma^2\theta^2\right] \\ &= \frac{1}{\sigma^2\tau^2}\left[(\sigma^2+\tau^2)\theta^2-2\tau^2 X_j\theta+\tau^2 X_j{}^2\right] \\ &= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left[\theta^2-2\frac{\tau^2}{\sigma^2+\tau^2}X_j\theta+\frac{\tau^2}{\sigma^2+\tau^2}X_j{}^2\right] \\ &= \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left[\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2\right. \\ &\qquad \left.\underbrace{-\left(\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2+\frac{\tau^2}{\sigma^2+\tau^2}X_j{}^2}\right] \end{aligned} \tag{23}$$

$$\begin{aligned} * \quad & \frac{\tau^2}{\sigma^2+\tau^2}X_j{}^2-\left(\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2 \\ = \quad & X_j{}^2\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)\left(1-\frac{\tau^2}{\sigma^2+\tau^2}\right) \\ = \quad & X_j{}^2\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right) \\ = \quad & X_j{}^2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)^2} \end{aligned} \tag{24}$$

Substituting (24) into (23) yields

$$\frac{(X_j-\theta)^2}{\sigma^2}+\frac{\theta^2}{\tau^2}=$$

$$=\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2+\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}X_j{}^2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)^2}$$

$$=\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2+\frac{X_j{}^2}{\sigma^2+\tau^2} \qquad (25)$$

Now use the completed square in (25) to continue the computation in (22).

$$f(X_j)$$

$$=\frac{1}{2\pi\sigma\tau}\int_{-\infty}^{\infty}e^{-\frac{1}{2}\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)^2}e^{-\frac{1}{2}\frac{X_j{}^2}{\sigma^2+\tau^2}}d\theta$$

$$=\frac{e^{-\frac{X_j{}^2}{2(\sigma^2+\tau^2)}}}{2\pi\sigma\tau}\int_{-\infty}^{\infty}\exp\left[-\left(\frac{\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j}{\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}}\right)^2\right]d\theta$$

Then

$$d\varphi = d\theta\Big/\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}$$

$$d\theta = \sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\,d\varphi$$

$$\theta=\pm\infty \implies \varphi=\pm\infty$$

$$\varphi=\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X_j\right)\Big/\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}$$

5

and

$$f(X_j) = \frac{e^{-\frac{X_j^2}{2(\sigma^2+\tau^2)}}}{2\pi\sigma\tau}\int_{-\infty}^{\infty} e^{-\varphi^2}\sqrt{\frac{2\sigma^2\tau^2}{\sigma^2+\tau^2}}\,d\varphi$$

$$= \frac{e^{-\frac{X_j^2}{2(\sigma^2+\tau^2)}}}{\pi\sqrt{2(\sigma^2+\tau^2)}}\underbrace{\int_{-\infty}^{\infty} e^{-\varphi^2}\,d\varphi}_{\sqrt{\pi}}$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\,e^{-\frac{X_j^2}{2(\sigma^2+\tau^2)}}$$

Therefore

$$f(X_j) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\,e^{-\frac{X_j^2}{2(\sigma^2+\tau^2)}} \qquad (26)$$

A.3 Calculation of the posterior distribution of θ_j

Since the subscript j remains constant throughout the calculation, it will be dropped in this appendix. Herein, θ_j will be replaced by θ, and X_j by X.

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{f(X)} = \frac{f(X,\theta)}{f(X)}$$

$$= \frac{(1/2\pi\sigma\tau)\ \exp\left[-\left(\frac{\theta^2}{2\tau^2}+\frac{(X-\theta)^2}{2\sigma^2}\right)\right]}{\left(1/\sqrt{2\pi(\sigma^2+\tau^2)}\right)\ \exp\left[-\frac{X^2}{2(\sigma^2+\tau^2)}\right]}$$

$$= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}\ \exp\left[-\frac{1}{2}\underbrace{\left(\frac{\theta^2}{\tau^2}+\frac{(X-\theta)^2}{\sigma^2}-\frac{X^2}{\sigma^2+\tau^2}\right)}\right]$$

- $\frac{\theta^2}{\tau^2}+\frac{(X-\theta)^2}{\sigma^2}-\frac{X^2}{\sigma^2+\tau^2} =$

$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\sigma^2(\sigma^2+\tau^2)\theta^2 + \tau^2(\sigma^2+\tau^2)\overbrace{(X-\theta)^2}^{X^2-2X\theta+\theta^2} - \sigma^2\tau^2X^2\Big]$$

$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\theta^2\left(\sigma^2(\sigma^2+\tau^2)+\tau^2(\sigma^2+\tau^2)\right) - 2\tau^2(\sigma^2+\tau^2)X\theta + X^2\left(\tau^2(\sigma^2+\tau^2)-\sigma^2\tau^2\right)\Big]$$

$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\Big[\theta^2(\sigma^2+\tau^2)^2 - 2(\sigma^2+\tau^2)\theta\cdot\tau^2X+\tau^4X^2\Big]$$

$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}\left((\sigma^2+\tau^2)\theta-\tau^2X\right)^2$$

$$= \frac{1}{\sigma^2\tau^2(\sigma^2+\tau^2)}(\sigma^2+\tau^2)^2\left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X\right)^2$$

$$= \left(\theta-\frac{\tau^2}{\sigma^2+\tau^2}X\right)^2 \Big/ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$$

Therefore,

$$\pi(\theta|X) = \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}} \exp\left[-\frac{\left(\theta - \frac{\tau^2}{\sigma^2+\tau^2}X\right)^2}{2\left(\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)}\right] \qquad (27)$$

A.4 Proof of the fact that n independent observations from the Normal population $\mathcal{N}(\theta, \sigma^2)$ can be treated as a single observation from $\mathcal{N}(\theta, \sigma^2/n)$

Given the data y, $f(y|\theta)$ can be viewed as a function of θ. We then call it the *likelihood function* of θ for given y, and writes

$$l(\theta|y) \propto f(y|\theta).$$

When y is a single data point from $\mathcal{N}(\theta, \sigma^2)$,

$$l(\theta|y) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta - x}{\sigma}\right)^2\right] = \exp\left[-\frac{1}{2\sigma^2}(\theta - x)^2\right], \tag{28}$$

where x is some function of y.

Now, suppose that $\vec{y} = (y_1, \ldots, y_n)$ represents a vector of n independent observations from $\mathcal{N}(\theta, \sigma^2)$. We can denote the sample mean by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The likelihood function of θ given such n independent observations from $\mathcal{N}(\theta, \sigma^2)$ is

$$l(\theta|\vec{y}) \propto \prod_i \exp\left[-\frac{1}{2\sigma^2}(y_i - \theta)^2\right] = \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right].$$

Also, since

$$\sum_{i=1}^{n}(y_i - \theta)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \theta)^2, \tag{29}$$

it follows that

$$\begin{aligned} l(\theta|\vec{y}) &\propto \underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_i (y_i - \bar{y})^2\right]}_{\text{const w.r.t. } \theta} \exp\left[-\frac{1}{2\sigma^2}n(\bar{y} - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{y})^2\right], \end{aligned} \tag{30}$$

which is a Normal function with mean $\bar{y}$ and variance σ^2/n. Comparing with (28), we can recognize that this is equivalent to treating the data $\vec{y}$ as a single observation $\bar{y}$ with mean θ and variance σ^2/n, i.e.,

$$\bar{y} \sim \mathcal{N}(\theta, \sigma^2/n). \tag{31}$$

PROOF OF (29):

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\theta)^2 &= \sum_i (y_i-\bar{y}+\bar{y}-\theta)^2 \\
&= \sum_i \left[(y_i-\bar{y})^2+2(y_i-\bar{y})(\bar{y}-\theta)+(\bar{y}-\theta)^2\right] \\
&= \sum_i (y_i-\bar{y})^2+2(\bar{y}-\theta)\sum_i (y_i-\bar{y})+\sum_i (\bar{y}-\theta)^2 \\
&= \sum_i (y_i-\bar{y})^2+2(\bar{y}-\theta)\underbrace{\left(\sum_i y_i-\sum_i \bar{y}\right)}_{n\bar{y}-n\bar{y}=0}+n(\bar{y}-\theta)^2 \\
&= \sum_i (y_i-\bar{y})^2+n(\bar{y}-\theta)^2
\end{aligned}$$

5

10

A.5 Distribution of the Sum of two Independent Normal Random Variables

Let

$$\begin{aligned} X &\sim \mathcal{N}(0,\alpha^2) \\ Y &\sim \mathcal{N}(0,\beta^2) \end{aligned}$$

be two independent random variables.

Claim: $X+Y \sim \mathcal{N}(0,\alpha^2+\beta^2)$
(This result is used for mean-0 Normal r.v.'s, although a more general result can be proven.)

Proof: (use moment generating functions)

$$\begin{aligned} m_X(t) &= \mathrm{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{1}{2\alpha^2}(x-0)^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}\alpha}\int_{-\infty}^{\infty} e^{-\frac{1}{2\alpha^2}\underbrace{[x^2-2\alpha^2 tx]}} dx \end{aligned} \tag{32}$$

Completing the square, we obtain

$$\begin{aligned} x^2-2\alpha^2 tx &= x^2-2(\alpha^2 t)x+(\alpha^2 t)^2-(\alpha^2 t)^2 \\ &= (x-\alpha^2 t)^2-(\alpha^4 t^2) \\ \frac{1}{\alpha^2}(x^2-2\alpha^2 tx) &= \left((x-\alpha^2 t)/\alpha\right)^2-(\alpha^4 t^2)/\alpha^2 \\ &= \left(\frac{x-\alpha^2 t}{\alpha}\right)^2-\alpha^2 t^2 \end{aligned} \tag{33}$$

Using the result of (33) in (32) yields

$$m_X(t) = \frac{e^{-\frac{1}{2}\left(-\alpha^2 t^2\right)}}{\sqrt{2\pi}\alpha}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\alpha^2 t}{\alpha}\right)^2} dx$$

$$\begin{aligned} \text{Let } y &= \frac{x-\alpha^2 t}{\alpha} \\ dy &= \frac{dx}{\alpha} \Longrightarrow dx = \alpha\, dy \end{aligned}$$

With this substitution, we obtain

$$m_X(t) = \frac{e^{\frac{1}{2}\alpha^2 t^2}}{\sqrt{2\pi}\alpha} \cdot \alpha \underbrace{\int_{y=-\alpha}^{\infty} e^{-\frac{1}{2}y^2} dy}_{\sqrt{2\pi}}$$

or

$$m_X(t) = e^{\frac{1}{2}\alpha^2 t^2} \tag{34}$$

Similarly

$$m_Y(t) = e^{\frac{1}{2}\beta^2 t^2} \tag{35}$$

To obtain the distribution of $X+Y$, it suffices to compute the corresponding moment generating function:

$$\begin{aligned} m_{X+Y}(t) &= \mathbf{E}\left(e^{t(X+Y)}\right) = \mathbf{E}\left(e^{tX}\, e^{tY}\right) \\ &= \mathbf{E}\left(e^{tX}\right)\mathbf{E}\left(e^{tY}\right) \quad \text{by independence of } X \text{ and } Y \\ &= m_X(t) \cdot m_Y(t) \\ &= e^{\frac{1}{2}\alpha^2 t^2} \cdot e^{\frac{1}{2}\beta^2 t^2} \quad \text{by (34) and (35)} \\ &= e^{\frac{1}{2}(\alpha^2+\beta^2)t^2}, \end{aligned}$$

which is a moment generating function of a Normal random variable with mean 0 and variance $\alpha^2+\beta^2$. Therefore,

$$X+Y \sim \mathcal{N}(0, \alpha^2+\beta^2). \tag{36}$$

5

A.6 Properties of the Chi-square Distribution

Let $X_1, X_2, \ldots, X_k$ be i.i.d.r.v.'s from standard Normal distribution, i.e.,

$$X_j \sim \mathcal{N}(0,1) \quad \forall j.$$

Then

$$\chi_k^2 = X_1^2 + X_2^2 + \cdots + X_k^2 = \sum_{j=1}^{k} X_j{}^2$$

is a random variable from Chi-square distribution with k degrees of freedom, denoted

$$\chi_k^2 \sim \chi_{(k)}^2.$$

It has the probability density function

$$f(x) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt. \tag{37}$$

The result we are using is

$$\mathbf{E}\left(\frac{1}{\chi_k^2}\right) = \frac{1}{k-2} \quad \text{for } k > 2,$$

which can be obtained as follows:

$$\begin{aligned} \mathbf{E}\left(\frac{1}{\chi_k^2}\right) &= \int_{\mathcal{R}} \frac{1}{x} f(x)\, dx \\ &= \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty \frac{1}{x} x^{k/2-1} e^{-x/2}\, dx \\ &= \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty x^{k/2-2} e^{-x/2}\, dx \end{aligned} \tag{38}$$

Let

$$\begin{array}{rclcrcl} t & = & x/2 & \Longrightarrow & x & = & 2t \\ & & & & dx & = & 2dt \\ x & = & 0 & \Longrightarrow & t & = & 0 \\ x & = & \infty & \Longrightarrow & t & = & \infty \end{array}$$

$$\begin{aligned} &\int_0^\infty x^{k/2-2} e^{-x/2}\, dx \\ &= \int_{t=0}^{\infty} (2t)^{k/2-2} e^{-t}\, 2\, dt \\ &= 2^{k/2-2} \cdot 2 \int_0^\infty t^{k/2-2} e^{-t}\, dt. \end{aligned} \tag{39}$$

Let

$$\begin{array}{rclcrcl} u & = & e^{-t} & & dv & = & t^{k/2-2}\, dt \\ du & = & -e^{-t}\, dt & & v & = & \frac{t^{k/2-1}}{k/2-1} \qquad \text{for } k > 2 \end{array}$$

Integration by parts transforms (39) into

$$\begin{aligned} &= 2^{k/2-1} \left(\frac{1}{k/2-1} \underbrace{e^{-t} t^{k/2-1} \Big|_0^\infty}_{\longrightarrow 0} - \int_0^\infty \frac{1}{k/2-1} t^{k/2-1} \left(-e^{-t}\right)\, dt \right) \\ &= \frac{2^{k/2-1}}{k/2-1} \underbrace{\int_0^\infty t^{k/2-1} e^{-t}\, dt}_{\Gamma(k/2),\ \text{by (37)}} \\ &= \frac{2^{k/2-1}}{k/2-1} \Gamma(k/2) \end{aligned}$$

Substituting this result in (38) yields

$$\begin{aligned} \mathbf{E}\left(\frac{1}{\chi_k^2}\right) &= \frac{1}{2^{k/2}\Gamma(k/2)} \cdot \frac{2^{k/2-1}\Gamma(k/2)}{k/2-1} \\ &= \frac{1}{2\,(k/2-1)} \\ &= \frac{1}{k-2} \quad \text{for } k > 2. \end{aligned} \tag{40}$$

A.7 Distribution of Sample Variance s^2

Let $X_j \sim \mathcal{N}(\mu, \sigma^2)$ for $j = 1, \ldots, n$ be independent r.v.'s. We'll derive the joint distribution of

$$\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma} \quad \text{and} \quad \frac{(n-1)\,s^2}{\sigma^2}.$$

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2 \\ \frac{(n-1)\,s^2}{\sigma^2} &= \frac{n-1}{\sigma^2} \cdot \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2 \\ &= \sum_{j=1}^{n} \left(\frac{X_j - \bar{X}}{\sigma}\right)^2 \end{aligned}$$

W.L.O.G. can reduce the problem to the case $\mathcal{N}(0,1)$, i.e., $\mu = 0$, $\sigma^2 = 1$: Let $Z_j = (X_j - \mu)/\sigma$. Then

$$\begin{aligned} \bar{Z} &= \frac{1}{n} \sum Z_j = \frac{1}{n} \sum \left(\frac{X_j - \mu}{\sigma}\right) = \frac{1}{n} \left(\frac{\sum X_j}{\sigma} - \frac{\sum \mu}{\sigma}\right) \\ &= \frac{1}{n} \left(\frac{\sum X_j}{\sigma} - \frac{n\mu}{\sigma}\right) = \frac{1}{\sigma} \left(\frac{\sum X_j}{n} - \mu\right) = \frac{\bar{X} - \mu}{\sigma} \end{aligned}$$

and hence

$$\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma} = \sqrt{n}\,\bar{Z}. \tag{41}$$

Also,

$$\begin{aligned} \frac{(n-1)\,s^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum \left(X_j - \bar{X}\right)^2 \\ &= \frac{1}{\sigma^2} \sum \left((X_j - \mu) + (\mu - \bar{X})\right)^2 \\ &= \sum \left[\frac{X_j - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma}\right]^2 = \sum \left(Z_j - \bar{Z}\right)^2 \end{aligned} \tag{42}$$

By (41) and (42), it suffices to derive the joint distribution of $\sqrt{n}\,\bar{Z}$ and $\sum_{j=1}^{n}\left(Z_j-\bar{Z}\right)^2$, where $Z_1,\ldots,Z_n$ are i.i.d. from $\mathcal{N}(0,1)$.

Let

$$P=\begin{pmatrix} \text{---}\,p_1\,\text{---} \\ \text{---}\,p_2\,\text{---} \\ \vdots \\ \text{---}\,p_n\,\text{---} \end{pmatrix}$$

be an $n\times n$ orthogonal matrix where

$$p_1=\left(\frac{1}{\sqrt{n}},\ldots,\frac{1}{\sqrt{n}}\right)$$

and the remaining rows p_j are obtained by, say, applying Gramm-Schmidt to $\{p_1,e_2,e_3,\ldots,e_n\}$, where e_j is a standard unit vector in j^{th} direction in $\mathcal{R}^n$. Let

$$\begin{aligned}\vec{Y}&=P\vec{Z}\\ &=\left(\begin{array}{cccc} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \hline & & \vdots & \\ \hline \end{array}\right)\begin{pmatrix} Z_1\\ Z_2\\ \vdots\\ Z_n\end{pmatrix}=\begin{pmatrix} Y_1\\ Y_2\\ \vdots\\ Y_n\end{pmatrix}\end{aligned}$$

Then

$$Y_1=\frac{1}{\sqrt{n}}\left(\sum_{j=1}^{n}Z_j\right)=\frac{1}{\sqrt{n}}n\bar{Z}=\sqrt{n}\,\bar{Z}. \tag{43}$$

Since P is orthogonal, it preserves vector lengths:

$$\begin{aligned}\|\vec{Y}\|^2 &= \|\vec{Z}\|^2\\ \sum_{j=1}^{n}{Y_j}^2 &= \sum_{j=1}^{n}{Z_j}^2\\ \Longrightarrow\left(\sum_{j=1}^{n}{Y_j}^2\right)-Y_1^2 &= \sum_{j=1}^{n}{Z_j}^2-\left(\sqrt{n}\,\bar{Z}\right)^2 \quad \text{by (43)}\end{aligned}$$

Hence

$$\begin{aligned}\sum_{j=2}^{n} Y_j{}^2 &= \sum_{j=1}^{n} Z_j{}^2 - n\bar{Z}^2 = \sum_{j=1}^{n} Z_j{}^2 - 2n\bar{Z}^2 + n\bar{Z}^2 \\ &= \sum_{j=1}^{n} Z_j{}^2 - 2\bar{Z}(n\bar{Z}) + n\bar{Z}^2 \\ &= \sum_{j=1}^{n} Z_j{}^2 - 2\bar{Z}\left(\sum_{j=1}^{n} Z_j\right) + \sum_{j=1}^{n} \bar{Z}^2 \\ &= \sum_{j=1}^{n} (Z_j - \bar{Z})^2 \qquad (44)\end{aligned}$$

Since the Y_j's are mutually independent (by orthogonality of P), we can conclude that

$$\sum_{j=2}^{n} Y_j{}^2 = \sum_{j=1}^{n} (Z_j - \bar{Z})^2$$

is independent of

$$Y_1 = \sqrt{n}\,\bar{Z}.$$

Also by orthogonality of P, $Y_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, n$, so

$$\left(\sum_{j=2}^{n} Y_j{}^2\right) \sim \chi^2_{(n-1)} \quad \text{(See Appendix A.6)}$$

and hence, by (42) and (44),

$$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{(n-1)} \qquad (45)$$

Since $\mathbf{E}\left(\chi^2_k\right) = k$, for $\chi^2_k \sim \chi^2_{(k)}$, we can see that

$$\mathbf{E}\left(\frac{(n-1)\,s^2}{\sigma^2}\right) = n - 1.$$

5

Also, since

$$\mathbf{E}\left(\frac{(n-1)\,s^2}{\sigma^2}\right) = \frac{n-1}{\sigma^2}\,\mathbf{E}\left(s^2\right),$$

we can conclude that

$$\mathbf{E}\left(s^2\right) = \frac{\sigma^2}{n-1}\cdot\frac{n-1}{\sigma^2}\,\mathbf{E}\left(s^2\right) = \frac{\sigma^2}{n-1}\cdot(n-1) = \sigma^2, \qquad (46)$$

i.e., s^2 is an unbiased estimator of the variance σ^2.

Various publications have been referenced herein, the contents of which are hereby incorporated by reference in their entireties. It should be noted that all procedures and algorithms according to the present invention described herein can be performed using the exemplary systems of the present invention illustrated in Figures 1 and 2 and described herein, as well as being programmed as software arrangements according to the present invention to be executed by such systems or other exemplary systems and/or processing arrangements.

WHAT IS CLAIMED IS:

1. A method for determining an association between a first dataset and a second dataset comprising:

a) obtaining at least one first data corresponding to one or more prior assumptions regarding said first and second datasets;

b) obtaining at least one second data corresponding to one or more portions of actual information regarding said first and second datasets; and

c) combining the at least one first data and the at least one second data to determine the association between the first and second datasets.

2. The method of Claim 1, wherein one of the one or more prior assumptions is that the means of the first and second datasets are random variables with a known *a priori* distribution.

3. The method of Claim 1, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ which is a mean, and τ^2 which is a variance may be unknown.

4. The method of Claim 1, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ is known.

5. The method of Claim 1, wherein one of the one or more prior assumptions is that the means of the first and second datasets are zero-mean normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein $\mu=0$.

6. The method of Claim 1, wherein one of the one or more portions of the actual information is an *a posteriori* distribution of the means of the first and second datasets obtained directly from the first and second datasets.

7. The method of Claim 1, wherein the association is a correlation.

8. The method of Claim 1, wherein the association is a dot product.

9. The method of Claim 1, wherein the association is a Euclidean distance.

10. The method of Claim 7, wherein the determination of the correlation comprises a use of James-Stein Shrinkage estimators in conjunction with the first and second data.

11. The method of Claim 10, wherein the determination of the correlation utilizes a correlation coefficient that is modified by an optimal shrinkage parameter γ.

12. The method of Claim 11, wherein determination of the optimal shrinkage parameter γ comprises the use of Bayesian considerations in conjunction with the first and second data.

13. The method of Claim 11, wherein the shrinkage parameter γ is estimated from the datasets using cross-validation.

14. The method of Claim 11, wherein the shrinkage parameter γ is estimated by simulation.

15. The method of Claim 11, wherein the correlation coefficient includes a plurality of correlation coefficients parameterized by $0 \leq \gamma \leq 1$ and may be defined, for datasets X_j and X_k as:

$$S(X_j, X_k) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k}\right),$$

wherein

$$\Phi_j{}^2 = \frac{1}{N}\sum_{i}\left(X_{ij} - (X_j)_{offset}\right)^2$$

16. The method of Claim 15, wherein $\gamma = \underbrace{\left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik} - Y_k)^2}{\sum_{k=1}^{M} Y_k{}^2}\right)}_{\gamma} Y_j$,

wherein *M* represents, in an M x N matrix, a number of rows corresponding to datapoints from the first dataset, and *N* represents a number of columns corresponding to datapoints from the second dataset.

17. The method of Claim 16, wherein *M* is the number of rows corresponding to all genes from which expression data has been collected in one or more microarray experiments.

18. The method of Claim 16, wherein *M* is representative of a genotype and *N* is representative of a phenotype.

19. The method of Claim 18, wherein the correlation is a genotype/phenotype correlation.

20. The method of Claim 19, wherein the genotype/phenotype correlation is indicative of a causal relationship between a genotype and a phenotype.

21. The method of Claim 20, wherein the phenotype is that of a complex genetic disorder.

22. The method of Claim 21, wherein the complex genetic disorder includes at least one of a cancer, a neurological disease, a developmental disorder, a neurodevelopmental disorder, a cardiovascular disease, a metabolic disease, an immunologic disorder, an infectious disease, and an endocrine disorder.

23. The method of Claim 7 wherein the correlation is provided between financial information for one or more financial instruments traded on a financial exchange.

24. The method of Claim 7 wherein the correlation is provided between user profiles for one or more users in an e-commerce application.

25. A software arrangement which, when executed on a processing device, configures the processing device to determine an association between a first dataset and a second dataset, the software arrangement comprising a processing subsystem which, when executed on the processing device, configures the processing device to perform the following steps:

a) obtaining at least one first data corresponding to one or more prior assumptions regarding said first and second datasets;

b) obtaining at least one second data corresponding to one or more portions of actual information regarding said first and second datasets; and

c) combining the at least one first data and the at least one second data to determine the association between the first and second datasets.

26. The software arrangement of Claim 25, wherein one of the one or more prior assumptions is that the means of the first and second datasets are random variables with a known *a priori* distribution.

27. The software arrangement of Claim 25, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ which is a mean, and τ^2 which is a variance may be unknown.

28. The software arrangement of Claim 25, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ is known.

29. The software arrangement of Claim 25, wherein one of the one or more prior assumptions is that the means of the first and second datasets are zero-mean normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein $\mu=0$.

30. The software arrangement of Claim 25, wherein one of the one or more portions of the actual information is an *a posteriori* distribution of the means of the first and second datasets obtained directly from the first and second datasets.

31. The software arrangement of Claim 25, wherein the association is a correlation.

32. The software arrangement of Claim 25, wherein the association is a dot product.

33. The software arrangement of Claim 25, wherein the association is a Euclidean distance.

34. The software arrangement of Claim 31, wherein the determination of the correlation comprises a use of James-Stein Shrinkage estimators in conjunction with the first and second data.

35. The software arrangement of Claim 34, wherein the determination of the correlation utilizes a correlation coefficient that is modified by an optimal shrinkage parameter γ.

36. The software arrangement of Claim 35, wherein determination of the optimal shrinkage parameter γ comprises the use of Bayesian considerations in conjunction with the first and second data.

37. The software arrangement of Claim 35, wherein the shrinkage parameter γ is estimated from the datasets using cross-validation.

38. The software arrangement of Claim 35, wherein the shrinkage parameter γ is estimated by simulation.

39. The software arrangement of Claim 35, wherein the correlation coefficient includes a plurality of correlation coefficients parameterized by $0 \leq \gamma \leq 1$ and may be defined, for datasets X_j and X_k as:

$$S(X_j, X_k) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k}\right),$$

wherein $$\Phi_j{}^2 = \frac{1}{N}\sum_{i}\left(X_{ij} - (X_j)_{offset}\right)^2$$

40. The software arrangement of Claim 39, wherein γ

$$= \underbrace{\left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik} - Y_k)^2}{\sum_{k=1}^{M} Y_k{}^2}\right)}_{\gamma} Y_j$$

where M represents, in an M x N matrix, a number of rows corresponding to datapoints from the first dataset, and N represents a number of columns corresponding to datapoints from the second dataset.

41. The software arrangement of Claim 40, wherein M is the number of rows corresponding to all genes from which expression data has been collected in one or more microarray experiments.

42. The software arrangement of Claim 40, wherein M is representative of a genotype and N is representative of a phenotype.

43. The software arrangement of Claim 42, wherein the correlation is a genotype/phenotype correlation.

44. The software arrangement of Claim 43, wherein the genotype/phenotype correlation is indicative of a causal relationship between a genotype and a phenotype.

45. The software arrangement of Claim 44, wherein the phenotype is that of a complex genetic disorder.

46. The software arrangement of Claim 45, wherein the complex genetic disorder includes at least one of a cancer, a neurological disease, a developmental disorder, a neurodevelopmental disorder, a cardiovascular disease, a metabolic disease, an immunologic disorder, an infectious disease, and an endocrine disorder.

47. The software arrangement of Claim 31 wherein the correlation is provided between financial information for one or more financial instruments traded on a financial exchange.

48. The software arrangement of Claim 31 wherein the correlation is provided between user profiles for one or more users in an e-commerce application.

49. A storage medium which includes thereon a software arrangement for determining an association between a first dataset and a second dataset, the software arrangement comprising a processing subsystem which, when executed on the processing device, configures the processing device to perform the following steps:

a) obtaining at least one first data corresponding to one or more prior assumptions regarding said first and second datasets;

b) obtaining at least one second data corresponding to one or more portions of actual information regarding said first and second datasets; and

c) combining the at least one first data and the at least one second data to determine the association between the first and second datasets.

50. The storage medium of Claim 49, wherein one of the one or more prior assumptions is that the means of the first and second datasets are random variables with a known *a priori* distribution.

51. The storage medium of Claim 49, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ which is a mean, and τ^2 which is a variance may be unknown.

52. The storage medium of Claim 49, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein parameter μ is known.

53. The storage medium of Claim 49, wherein one of the one or more prior assumptions is that the means of the first and second datasets are zero-mean normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein $\mu=0$.

54. The storage medium of Claim 49, wherein one of the one or more portions of the actual information is an *a posteriori* distribution of the means of the first and second datasets obtained directly from the first and second datasets.

55. The storage medium of Claim 49, wherein the association is a correlation.

56. The storage medium of Claim 49, wherein the association is a dot product.

57. The storage medium of Claim 49, wherein the association is a Euclidean distance.

58. The storage medium of Claim 55, wherein the determination of the correlation comprises a use of James-Stein Shrinkage estimators in conjunction with the first and second data.

59. The storage medium of Claim 58, wherein the determination of the correlation utilizes a correlation coefficient that is modified by an optimal shrinkage parameter γ.

60. The storage medium of Claim 59, wherein determination of the optimal shrinkage parameter γ comprises the use of Bayesian considerations in conjunction with the first and second data.

61. The storage medium of Claim 59, wherein the shrinkage parameter γ is estimated from the datasets using cross-validation.

62. The storage medium of Claim 59, wherein the shrinkage parameter γ is estimated by simulation.

63. The storage medium of Claim 59, wherein the correlation coefficient includes a plurality of correlation coefficients parameterized by $0 \leq \gamma \leq 1$ and may be defined, for datasets X_j and X_k as:

$$S(X_j, X_k) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j} \right) \left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k} \right),$$

wherein

$$\Phi_j{}^2 = \frac{1}{N}\sum_i \left(X_{ij} - (X_j)_{offset}\right)^2 .$$

64. The storage medium of Claim 63, wherein γ

$$= \left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik} - Y_k)^2}{\sum_{k=1}^{M} Y_k{}^2}\right) Y_j$$

$$\gamma$$

where M represents, in an M x N matrix, a number of rows corresponding to datapoints from the first dataset, and N represents a number of columns corresponding to datapoints from the second dataset.

65. The storage medium of Claim 64, wherein M is the number of rows corresponding to all genes from which expression data has been collected in one or more microarray experiments.

66. The storage medium of Claim 64, wherein M is representative of a genotype and N is representative of a phenotype.

67. The storage medium of Claim 66, wherein the correlation is a genotype/phenotype correlation.

68. The storage medium of Claim 67, wherein the genotype/phenotype correlation is indicative of a causal relationship between a genotype and a phenotype.

69. The storage medium of Claim 68, wherein the phenotype is that of a complex genetic disorder.

70. The storage medium of Claim 69, wherein the complex genetic disorder includes at least one of a cancer, a neurological disease, a developmental disorder, a neurodevelopmental disorder, a cardiovascular disease, a metabolic disease, an immunologic disorder, an infectious disease, and an endocrine disorder.

71. The storage medium of Claim 55, wherein the correlation is provided between financial information for one or more financial instruments traded on a financial exchange.

72. The storage medium of Claim 55, wherein the correlation is provided between user profiles for one or more users in an e-commerce application.

73. A system for determining an association between a first dataset and a second dataset comprising:

a) obtaining at least one first data corresponding to one or more prior assumptions regarding said first and second datasets;

b) obtaining at least one second data corresponding to one or more portions of actual information regarding said first and second datasets; and

c) combining the at least one first data and the at least one second data to determine the association between the first and second datasets.

74. The system of Claim 73, wherein one of the one or more prior assumptions is that the means of the first and second datasets are random variables with a known *a priori* distribution.

75. The system of Claim 73, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ which is a mean, and τ^2 which is a variance may be unknown.

76. The system of Claim 73, wherein one of the one or more prior assumptions is that the means of the first and second datasets are normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein μ is known.

77. The system of Claim 73, wherein one of the one or more prior assumptions is that the means of the first and second datasets are zero-mean normal random variables with an *a priori* Gaussian distribution $N(\mu, \tau^2)$, wherein $\mu=0$.

78. The system of Claim 73, wherein one of the one or more portions of the actual information is an *a posteriori* distribution of the means of the first and second datasets obtained directly from the first and second datasets.

79. The system of Claim 73, wherein the association is a correlation.

80. The system of Claim 73, wherein the association is a dot product.

81. The system of Claim 73, wherein the association is a Euclidean distance.

82. The system of Claim 79, wherein the determination of the correlation comprises a use of James-Stein Shrinkage estimators in conjunction with the first and second data.

83. The system of Claim 82, wherein the determination of the correlation utilizes a correlation coefficient that is modified by an optimal shrinkage parameter γ.

84. The system of Claim 83, wherein determination of the optimal shrinkage parameter γ comprises the use of Bayesian considerations in conjunction with the first and second data.

85. The system of Claim 83, wherein the shrinkage parameter γ is estimated from the datasets using cross-validation.

86. The system of Claim 83, wherein the shrinkage parameter γ is estimated by simulation.

87. The system of Claim 83, wherein the correlation coefficient includes a plurality of correlation coefficients parameterized by $0 \leq \gamma \leq 1$ and may be defined, for datasets X_j and X_k as:

$$S(X_j, X_k) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j} \right) \left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k} \right),$$

wherein

$$\Phi_j{}^2 = \frac{1}{N}\sum_i \left(X_{ij} - (X_j)_{offset}\right)^2 \quad .$$

88. The system of Claim 87, wherein γ

$$= \underbrace{\left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik} - Y_k)^2}{\sum_{k=1}^{M} Y_k{}^2}\right) Y_j}_{\gamma}$$

where M represents, in an M x N matrix, a number of rows corresponding to datapoints from the first dataset, and N represents a number of columns corresponding to datapoints from the second dataset.

89. The system of Claim 88, wherein M is the number of rows corresponding to all genes from which expression data has been collected in one or more microarray experiments.

90. The system of Claim 88, wherein M is representative of a genotype and N is representative of a phenotype.

91. The system of Claim 90, wherein the correlation is a genotype/phenotype correlation.

92. The system of Claim 91, wherein the genotype/phenotype correlation is indicative of a causal relationship between a genotype and a phenotype.

93. The system of Claim 92, wherein the phenotype is that of a complex genetic disorder.

94. The system of Claim 93, wherein the complex genetic disorder includes at least one of a cancer, a neurological disease, a developmental disorder, a neurodevelopmental disorder, a cardiovascular disease, a metabolic disease, an immunologic disorder, an infectious disease, and an endocrine disorder.

95. The system of Claim 79, wherein the correlation is provided between financial information for one or more financial instruments traded on a financial exchange.

96. The system of Claim 79, wherein the correlation is provided between user profiles for one or more users in an e-commerce application.

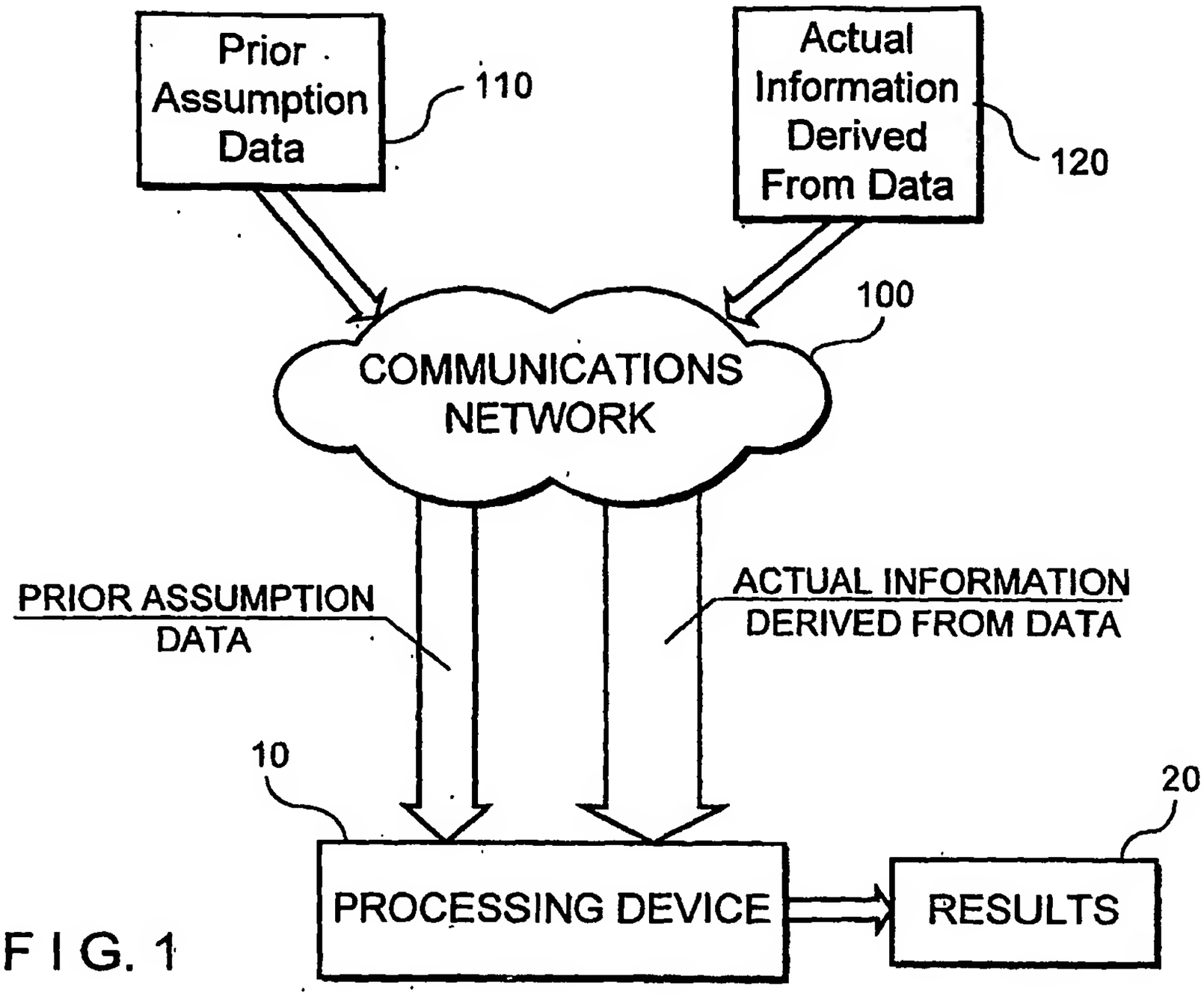

Prior Assumption Data
110
Actual Information Derived From Data
120
100
COMMUNICATIONS NETWORK
PRIOR ASSUMPTION DATA
ACTUAL INFORMATION DERIVED FROM DATA
10
20
PROCESSING DEVICE
RESULTS

FIG. 1

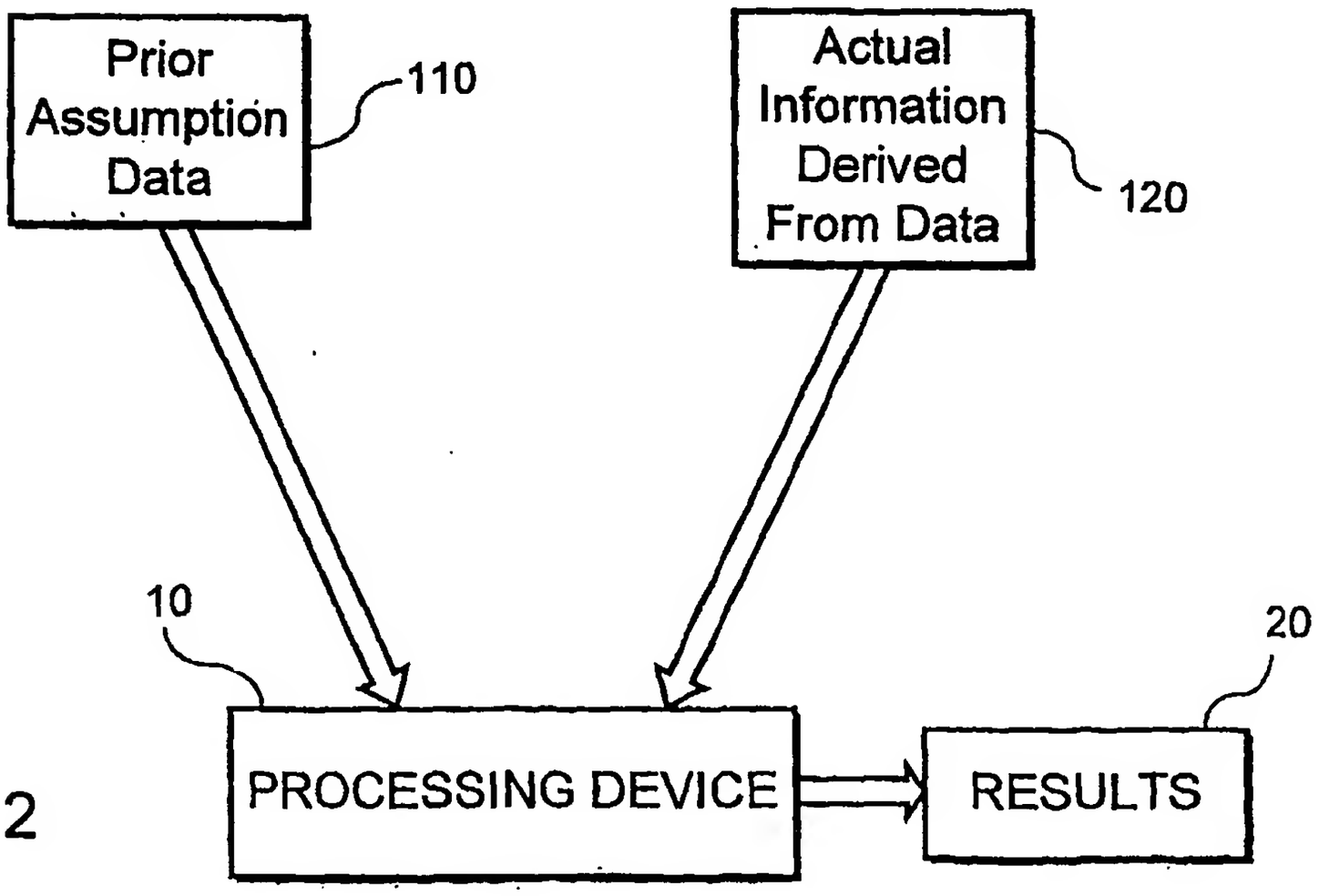

Prior Assumption Data
110
Actual Information Derived From Data
120
10
20
PROCESSING DEVICE
RESULTS

FIG. 2

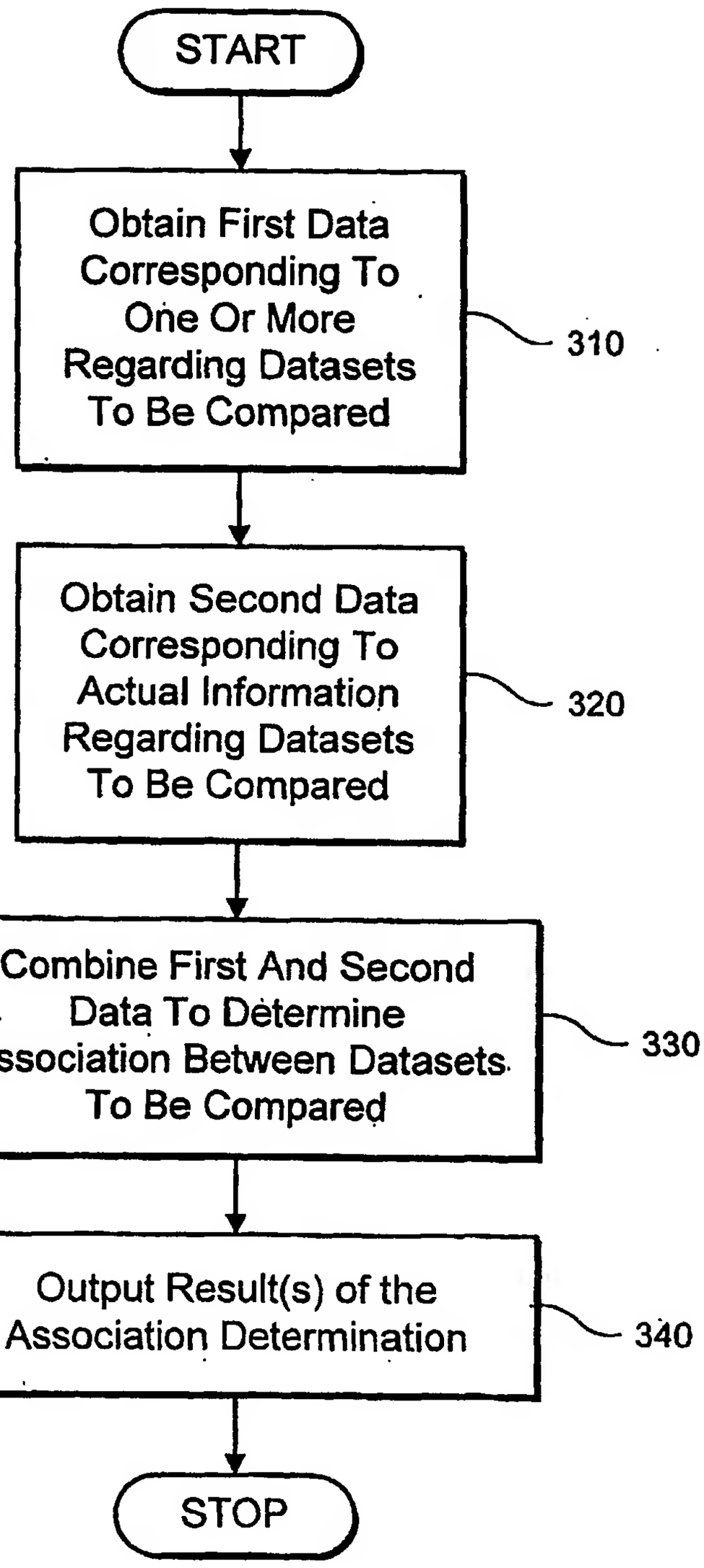

START
Obtain First Data Corresponding To One Or More Regarding Datasets To Be Compared
310
Obtain Second Data Corresponding To Actual Information Regarding Datasets To Be Compared
320
Combine First And Second Data To Determine Association Between Datasets To Be Compared
330
Output Result(s) of the Association Determination
340
STOP

FIG. 3

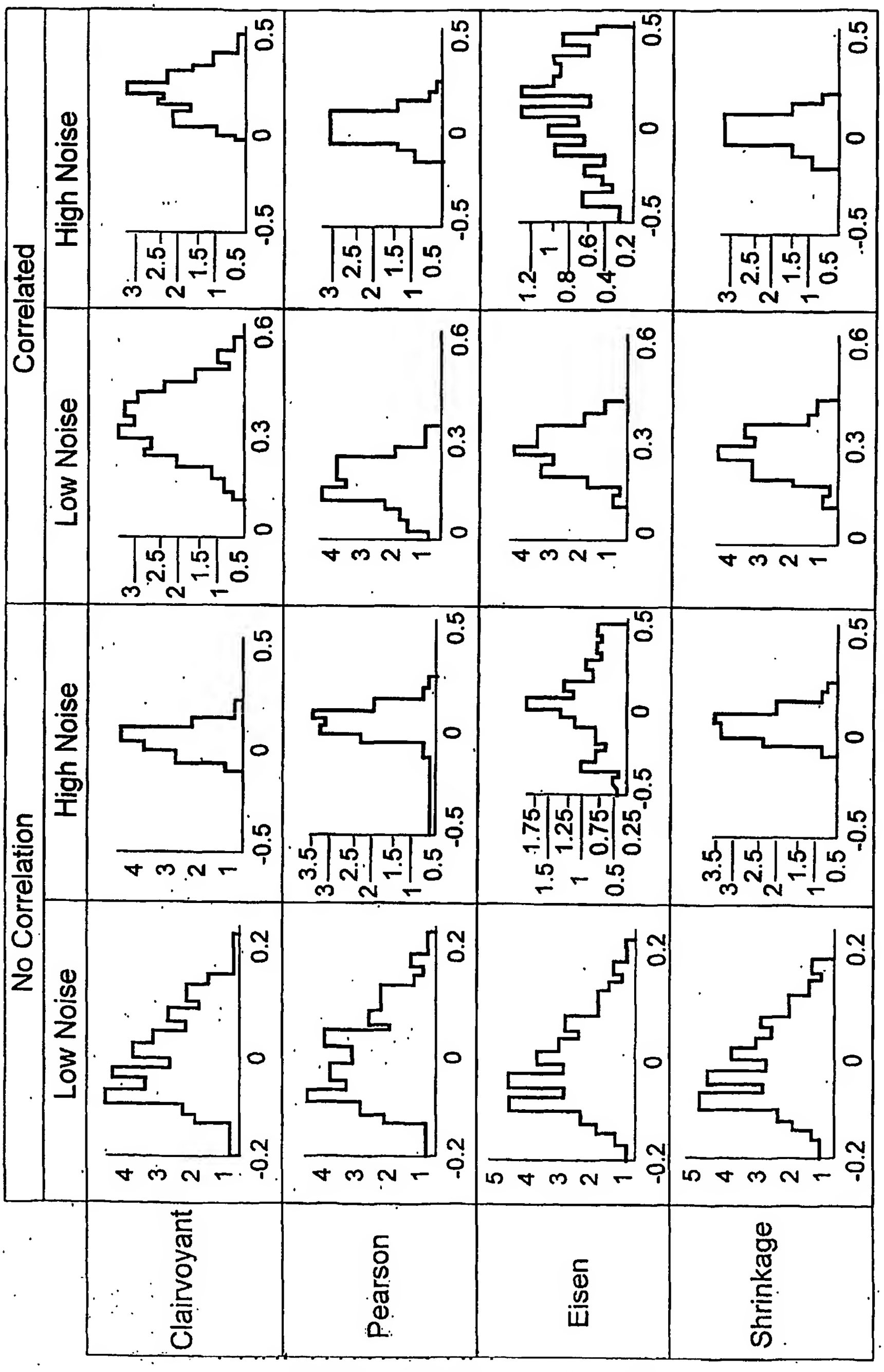

Correlated
Correlation
No Correlation
High Noise
Low Noise
High Noise
Low Noise
Clairvoyant
Pearson
Eisen
Shrinkage

FIG. 4

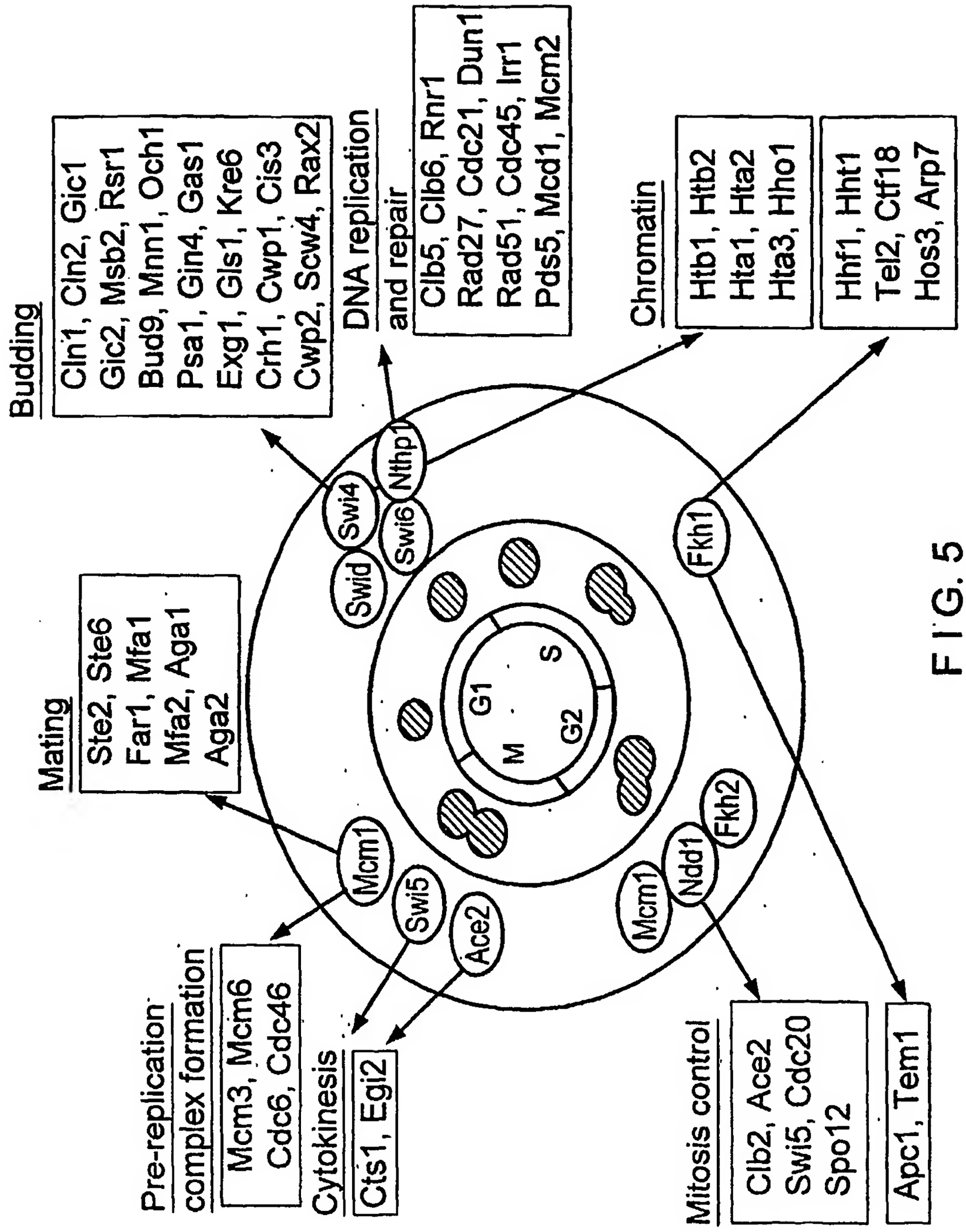

Budding
Cln1, Cln2, Gic1
Gic2, Msb2, Rsr1
Bud9, Mnn1, Och1
Psa1, Gin4, Gas1
Exg1, Gls1, Kre6
Crh1, Cwp1, Cis3
Cwp2, Scw4, Rax2
DNA replication and repair
Clb5, Clb6, Rnr1
Rad27, Cdc21, Dun1
Rad51, Cdc45, Irr1
Pds5, Mcd1, Mcm2
Chromatin
Htb1, Htb2
Hta1, Hta2
Hta3, Hho1
Hhf1, Hht1
Tel2, Ctf18
Hos3, Arp7
Mating
Ste2, Ste6
Far1, Mfa1
Mfa2, Aga1
Aga2
Swid
Swi4
Swi6
Nthp1
Fkh1
Mcm1
Swi5
Ace2
Mcm1
Ndd1
Fkh2
G1
S
G2
M
Pre-replication complex formation
Mcm3, Mcm6
Cdc6, Cdc46
Cytokinesis
Cts1, Egi2
Mitosis control
Clb2, Ace2
Swi5, Cdc20
Spo12
Apc1, Tem1

FIG. 5

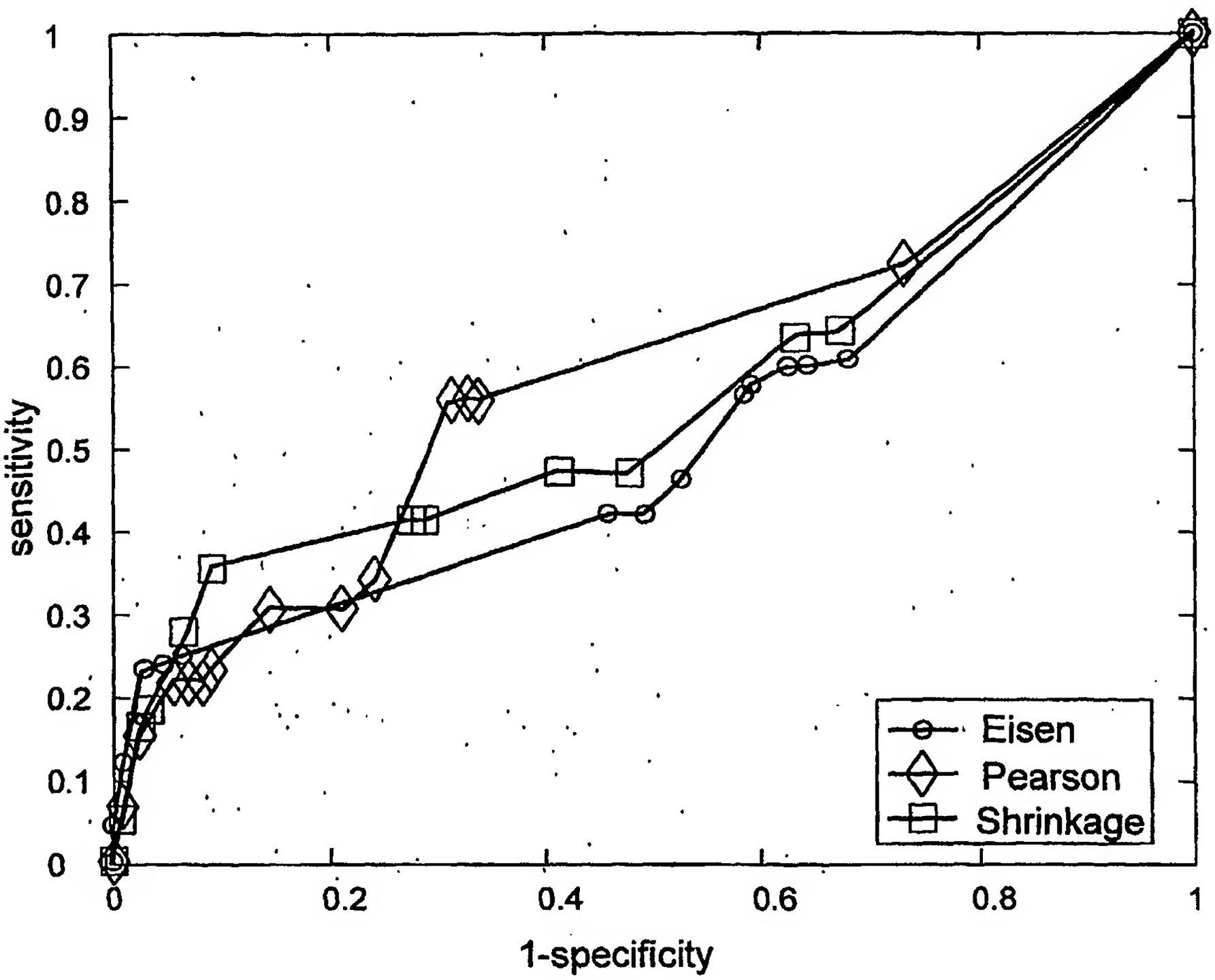

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
sensitivity
0
0.2
0.4
0.6
0.8
1
1-specificity
Eisen
Pearson
Shrinkage

FIG. 6

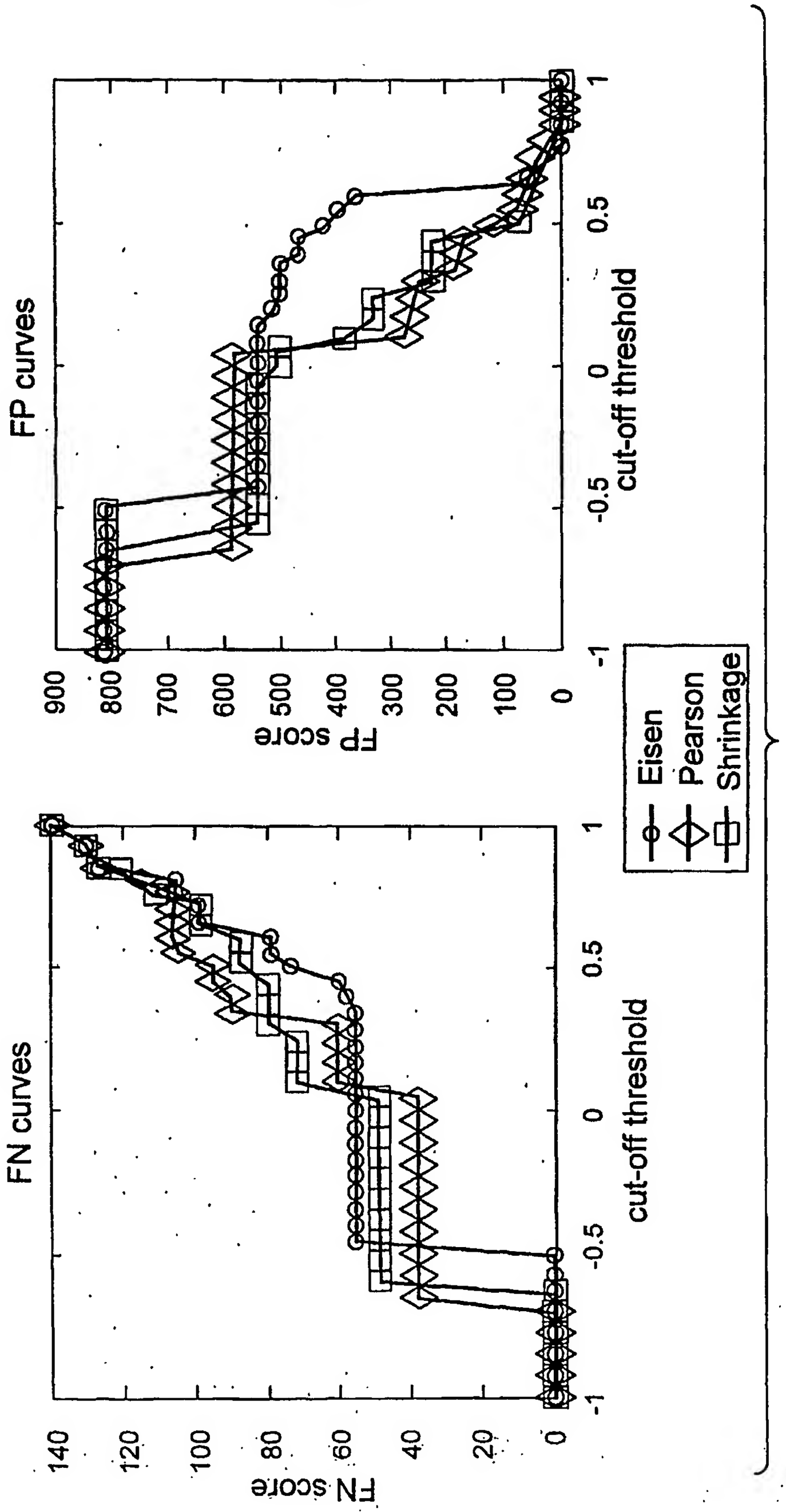

FP curves
FP score
900
800
700
600
500
400
300
200
100
0
1
0.5
0
-0.5
-1
cut-off threshold
Eisen
Pearson
Shrinkage
FN curves
FN score
140
120
100
80
60
40
20
0
1
0.5
0
-0.5
-1
cut-off threshold

FIG. 7

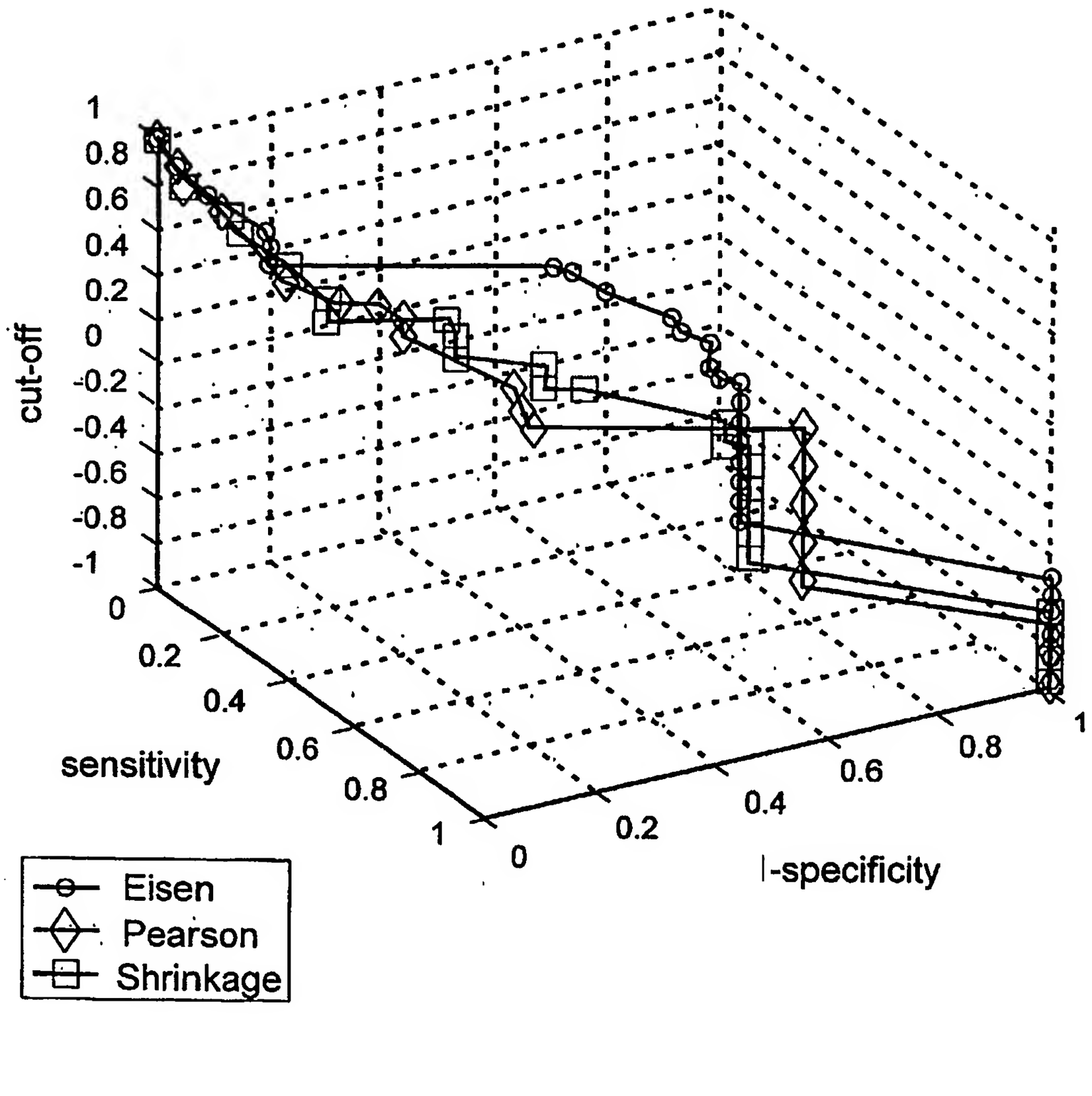

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
cut-off
0
0.2
0.4
0.6
0.8
1
sensitivity
0
0.2
0.4
0.6
0.8
1
1-specificity
Eisen
Pearson
Shrinkage

FIG. 8